How to interpret diagnostic tests

Ignacio Pérez[a], Iara Yamila Taito-Vicenti[b], Catalina Gracia González-Xuriguera[a], Cristhian Carvajal[a], Juan Víctor Ariel Franco[b], Cristóbal Loézar[c,d,*]

[a] Escuela de Medicina, Universidad de Valparaíso, Viña del Mar, Chile
[b] Instituto Universitario Hospital Italiano de Buenos Aires, Buenos Aires, Argentina
[c] Centro Interdisciplinario de Estudios en Salud (CIESAL), Universidad de Valparaíso, Viña del Mar, Chile
[d] Centro Asociado Universidad de Valparaíso, Cochrane Chile, Viña del Mar, Chile

*Corresponding author cristobal.loezar@uv.cl

Citation Pérez I, Taito-Vicenti IY, González-Xuriguera CG, Carvajal C, Franco JVA, Loézar C. How to interpret diagnostic tests. *Medwave* 2021;21(7):e8432

Doi 10.5867/medwave.2021.07.8432

Submission date 28/11/2020
Acceptance date 26/5/2021
Publication date 4/8/2021

Origin No solicitado

Type of review Con revisión por pares externa, por dos árbitros a doble ciego

Keywords diagnostic test, accuracy, impact, sensitivity, specificity

Abstract

Healthcare professionals make decisions in a context of uncertainty. When making a diagnosis, relevant patient characteristics are categorized to fit a particular condition that explains what the patient is experiencing. During the diagnostic process, tools such as the medical interview, physical examination, and other complementary tests support this categorization. These tools, known as diagnostic tests, allow professionals to estimate the probability of the presence or absence of the suspected medical condition. The usefulness of diagnostic tests varies for each clinical condition, and studies of accuracy (sensitivity and specificity) and diagnostic impact (impact on health outcomes) are used to evaluate them. In this article, the general theoretical and practical concepts about diagnostic tests in human beings are addressed, considering their historical background, their relationship with probability theories, and their practical utility with illustrative examples.

Main messages

- When deciding in uncertainty frameworks, practitioners do not have absolute certainty about a patient's diagnosed condition.
- Diagnostic tests support the diagnostic process in categorizing patient experiences into a particular medical condition that involves specific pathogenesis, treatment, and prognosis.
- The different diagnostic tests can range from questions in the anamnesis and signs on physical examination to complementary examinations (laboratory, imaging, or other procedures). These are assessed by accuracy and impact studies.
- This article offers a practical approach to the available reviews found in the main databases and specialized reference texts. It refers to tests and diagnostic accuracy in a friendly language, oriented to the training of undergraduate and graduate students.

Introduction

In the healthcare setting, professionals must make decisions in a context of uncertainty. In making a diagnosis, clinicians categorize patient experiences into a particular condition that involves specific pathogenesis, treatment, and prognosis[1]. However, in most cases there is no absolute certainty whether a patient has the diagnosed condition or not[2].

More than a century ago, diagnostics were based on anamnesis and physical examination. According to Erick Cobo and colleagues, the English monk Thomas Bayes concluded that God's existence could

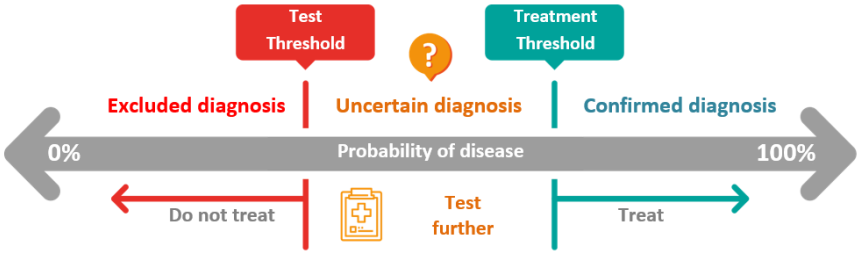
only be demonstrated if one first believed in God. Therefore, the probability that God exists depends on being a believer or not[3]. This reasoning applied to medical diagnosis states that the event probability after applying a test depends on the event probability prior to the test application, in addition to test characteristics[4]. The presumed probability prior to its test application is a process in which the health professional uses knowledge, experience, and clinical judgment[5].

In turn, there are other diagnostic approaches such as heuristics, defined by Perez[6] as "psychological mechanisms based on human performance in problem-solving, by which we reduce the uncertainty produced by our limitation in dealing with the complexity of environmental stimuli". Thus, it is a fast and intuitive way of thinking that provides probability estimates for decision-making. However, the use of heuristics carries potential avoidable errors that can lead to incorrect diagnoses[7] (Example 1). Evidence-based medicine provides tools to "objectify" clinical experience, avoid biases, and facilitate the interpretation of clinical scenarios.

Example 1.
A health professional suspects an irritable bowel syndrome diagnosis after examining a man with abdominal pain without alarm signs and general laboratory tests in a normal range. However, he saw a patient with similar clinical characteristics diagnosed with porphyria a week ago. For this reason, he decides to request specific tests to rule out this disease. This intuitive thinking corresponds to heuristic thinking, and it is based on recognizing familiar elements in new situations from recently remembered ("available") information.

The information provided by diagnostic methods increases or decreases the probability of a particular condition[8], moving in between the diagnostic threshold and the therapeutic threshold (Figure 1). The diagnostic threshold reflects the minimum probability needed to consider a particular condition plausible, whereas the therapeutic threshold reflects the confidence needed in the diagnosis to initiate treatment. Below the diagnostic threshold, testing is not worthwhile because the diagnostic probability is low[2,9]. Conversely, above the therapeutic threshold, the diagnosis has such a high probability that it justifies therapeutic decisions[2]. In between, when the diagnostic probability is intermediate, further testing is required to achieve a probability that is below the diagnostic threshold or above the treatment threshold[2,9].

Figure 1. Illustration of diagnostic and therapeutic thresholds.



In red is the diagnostic threshold that indicates the maximum probability that is tolerated to exclude a diagnosis. In green is the therapeutic threshold that indicates the minimum probability at which the diagnosis is assumed to be probable in order to initiate treatment, among other decisions. In orange is the intermediate probability at which further diagnostic tests are needed.
Source: Prepared by the authors.

Example 2.
A school-aged child with a low fever that started a few hours ago and without symptoms that could indicate a specific infectious focus has a low probability of having a urinary tract infection as a focus (below the diagnostic threshold). However, if the same patient also presents urinary symptoms, he/she would benefit from urine tests (between both thresholds). If the urine tests were compatible with a urinary tract infection, antibiotic treatment would be initiated (above the therapeutic threshold).

Diagnostic tests are a group of actions (including questions) to assess a patient's history, signs on physical examination, and complementary tests (laboratory, procedural, or imaging) to determine the presence or absence of a condition. In some cases, they are also used to establish its severity. Diagnostic tests are evaluated for accuracy and impact. Accuracy is defined as the probability that the test result correctly predicts the existence and absence of a particular condition. This can also be interpreted as the relative frequency of subjects in whom the test got the diagnosis right, represented by the following formula:

$$\frac{(\text{true positives} + \text{true negatives})}{(\text{total subjects evaluated})}$$

However, it is important to consider that a diagnostic test can be more accurate in identifying sick patients or identifying healthy individuals; therefore, it may be helpful or not depending on the specific scenario[9].

Also, the accuracy of diagnostic tests can be represented by indicators such as sensitivity, specificity, positive predictive value, negative predictive value, likelihood ratios, and receiver operating characteristic (ROC) curves. These indicators are usually familiar to most general practitioners. However, there is evidence that they can be misapplied[10].

Accuracy assessment is performed by comparing the results obtained from the diagnostic test evaluated with a reference standard in the same group of patients. The reference standard – also called gold standard – corresponds to a single test or combination of methods (composite gold standard), which allows establishing the presence or absence of a given condition[9]. For example, to diagnose

acute pulmonary thromboembolism, the gold standard is computed tomography angiography. If the D-dimer latex agglutination test were used to diagnose the same condition, the estimation of the sensitivity and specificity of the results would be from the comparison of these with the gold standard[11]. The impact of a diagnostic test refers to how much a given diagnostic test result impacts patient care[12]. Therefore, impact assessment determines how the information provided by the test result affects therapeutic decisions and clinical outcomes[13].

A prospective short- and long-term follow-up study should be performed to determine the impact of a diagnostic test. Another alternative is to perform a retrospective study that allows monitoring, among other things, the number of diagnostic tests applied and the time delay until a definitive diagnosis or a definitive treatment. As a practical illustration, a cerebral vascular accident, without therapeutic alternatives (surgical or endovascular) because cerebral images indicate a poor prognosis, knowing lesions characteristics through new diagnostic tests would not affect the patient's management[14].

This article is the seventh in a methodological series of thirteen narrative reviews on general topics in biostatistics and clinical epidemiology. This review explores and summarizes in a user-friendly language published articles available in the main databases and specialized reference texts. The series is oriented to the training of undergraduate and graduate students. It is carried out by the Chair of Evidence-Based Medicine of the School of Medicine of the Valparaíso University, Chile, in collaboration with the University Institute of the Italian Hospital of Buenos Aires, Argentina, and the UC Evidence Center of the Pontifical Catholic University of Chile. This manuscript aims to address the main theoretical and practical concepts of diagnostic testing in humans.

Probabilities and more probabilities in clinical reasoning

Probabilistic approaches are constantly used in medical practice to determine the probability that an individual has of suffering from a particular condition. This procedure is prior to performing a diagnostic test. This initial diagnostic approximation corresponds to the pre-test probability. This test depends on the clinician's subjective assessment of the presence or absence of semiology findings to diagnose a particular condition of interest[15,16]. In simplified form, it means that in the absence of additional relevant information, it has been accepted to use the condition's prevalence under study to estimate the pre-test probability[15].

A negative diagnostic test with a high clinical suspicion or high pre-test probability (Example 3), or a positive diagnostic test with a low pre-test probability (Example 4), will make us doubt the test result in the first instance. When the pre-test probability is intermediate, the diagnostic test result may modify the uncertain probabilistic scenario to rule out or confirm the diagnostic suspicion (Example 5).

Problems with testing in the absence of uncertainty

Example 3. High pre-test probability with a negative result.

A seven-year-old boy shows to the emergency department with odynophagia, fever higher than 38 degrees Celsius, with swollen and painful pultaceous and laterocervical lymphadenopathies. His mother reported that his ten-year-old brother had Streptococcus pyogenes pharyngitis (confirmed) less than five days ago. When a rapid test was performed, it was negative. Because the patient's pre-test probability is so high, one can consider the possibility of a false negative, i.e., that the test missed the presence of a disease. In this scenario, it would be appropriate to request the gold standard (pharyngeal culture) diagnostic test. It is important to note that if it is impossible to apply a diagnostic test, the therapeutic threshold could be lowered (Figure 1), and an "empirical" antibiotic treatment could be initiated.

Example 4. Low pre-test probability with a positive result.

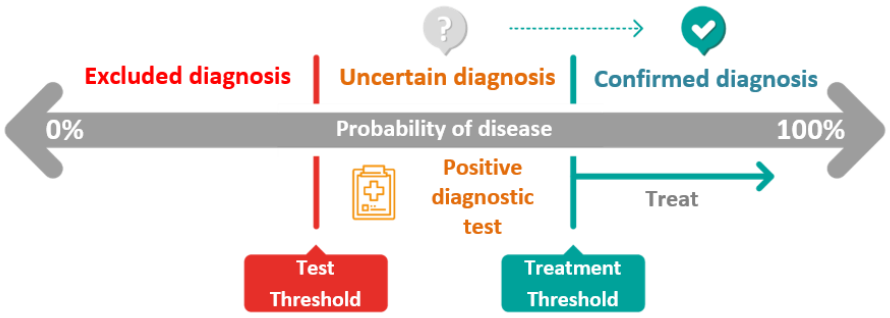
An 18-year-old healthy young man with a normal physical examination and no personal or family history of cardiovascular disease performs a graded ergometric test as part of the routine examinations before entering compulsory military service. During the test, the patient presents a horizontal ST-segment depression of 2 millimeters in DIII. As the patient's pre-test probability is very low, it is reasonable to think that the result is a false positive for acute myocardial infarction. This situation suggests that it is inappropriate to request a low specificity diagnostic test when the pre-test probability is very low, given that, in the event of a positive result, the patient should be subjected to more specific tests to confirm a false positive.

Tests in the area of uncertainty

Example 5. Intermediate pre-test probability with a positive result.

A 31-year-old female with no history of morbidities presents with three months of abdominal distension and colic pain associated with intermittent tenesmus and mucous diarrheal stools. Physical examination shows only hypogastric abdominal distension. She has a family history of inflammatory bowel disease and is a smoker. Given the diagnostic hypothesis of inflammatory bowel disease, fecal calprotectin was requested. The latter result was elevated, with a value of 150 micrograms per gram (the sensitivity and specificity for discriminating inflammatory bowel disease from irritable bowel syndrome vary according to the calprotectin cut-off point and ranges between 80 and 100% and between 74 and 100%, respectively)[17]. Since the probability after applying the test was elevated, ileocolonoscopy plus biopsy was requested, which showed cobblestone pattern, aphthous ulcers, mucosal fissures, and a biopsy compatible with Crohn's disease. Once the diagnosis was made, treatment was established.

Figure 2. Change in clinical behavior after applying a diagnostic test.



The patient in Example 5 starts in the diagnostic uncertainty zone because of his symptomatology and family history. When applying a fecal calprotectin diagnostic test, the elevated result increases the probability of diagnosing "inflammatory bowel disease" and exceeds the therapeutic threshold. Therefore, treatment should be initiated.
Source: Prepared by the authors.

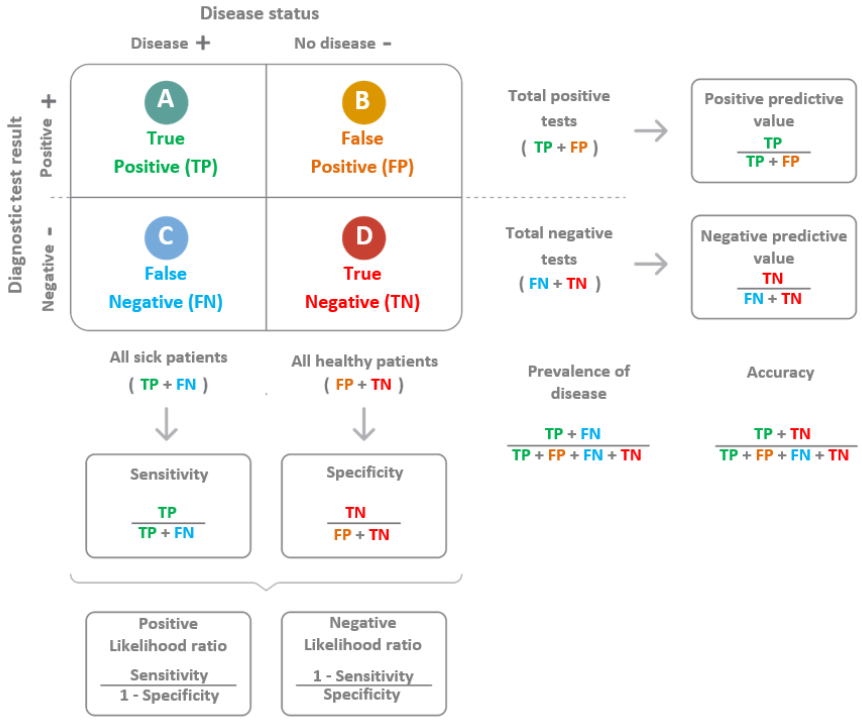
How do we measure diagnostic accuracy?

Sensitivity and specificity

When applying a diagnostic test, there is the possibility of incorrectly classifying individuals who have undergone the test. Examples are alleged sick people who are healthy (false positives) and alleged healthy people who are sick (false negatives). The information on the values obtained for the test, in contrast to the values of the reference test or gold standard, is presented in tabular format (Figure 3). The so-called "2x2 contingency tables" are constructed with two columns. According to the reference standard, the columns correspond to the positive (left) and negative (right) result of the condition. To these are added two rows reflecting the positive (top) or the negative (bottom) result of the condition, according to the index test. In addition, a letter is designated to each cell[9]:

A. True positives: those sick individuals with a positive test result.
B. False positives: those healthy individuals with a positive test result.
C. False negatives: those sick individuals with a negative test result.
D. True negatives: those healthy individuals with a negative test result.).

Figure 3. Contingency table for the estimation of diagnostic accuracy.



Source: Prepared by the authors.

Sensitivity" and "specificity"[2] are used to evaluate diagnostic tests. These are established values obtained from the diagnostic test application in a specific population at the validation time. In this sense, sensitivity and specificity are intrinsic properties of the diagnostic test. However, its performance also depends on the population characteristics in which it will be applied. These aspects are discussed in more detail later in the text[18].

Sensitivity is the probability that the test will correctly classify sick individuals or the probability that the sick individual will be positive[2]. Tests with high sensitivity are useful for screening because they have very few false negatives[19]. However, specificity is also important to avoid an excess of false positives, especially if these involve expensive or invasive confirmatory tests. In addition, because of the low number of false negatives, they are especially useful where failure to diagnose a specific disease or event may be dangerous or fatal to patients[16,18].

Example 7.

A 67-year-old woman presents with confusion, nausea, vomiting, and headache. A professional clinically evaluate her and suspects that she may have intracranial hypertension. As part of the neurological evaluation, he decides to perform a funduscopic examination, given that the loss of spontaneous retinal venous pulsation is a sign without false negatives for intracranial hypertension. Upon noting that pulsation is present, he considers the result as a true negative for intracranial hypertension.

Specificity is the probability that the test will correctly classify healthy individuals or the probability that healthy individuals will have a negative result[2]. A highly specific test has a very low false positive rate. Therefore, it has a high ability to confirm the disease. This means that if a highly specific test result is positive, there is a high chance of a true positive[18]. In clinical practice, tests with high specificity are preferred in confirming a diagnosis because of their low number of false positives. This is particularly important in severe diseases because timely mannered treatment can significantly reduce the physical, economic, and psychological consequences[16].

Example 8.

A 27-year-old female patient with a family history of Wilson's disease presents for consultation. Her physician seeks to evaluate the presence of the Kayser-Fleisher ring (golden rings in the Descemet membrane of the limbic region of the cornea) on physical examination. This sign is pathognomonic, i.e., it has a specificity of 100%. If this ring were present, it could be interpreted as confirmation of the disease since the high specificity suggests that false positives are not likely.

The estimation of the sensitivity and specificity of a diagnostic test will have greater applicability the broader the demographic and/or clinical characteristics of the sample of sick and healthy individuals in the population where the test will be used. Suppose the sample is representative of a population and the estimates are used in another population with different characteristics. In that case, the sensitivity and specificity values are incorrect, or at least not applicable to the population where the test is being used.

Since it is required to know patients' health/sick status to calculate the sensitivity and specificity, it is necessary to contrast the diagnosis using a method that proposes an ideal parameter or gold standard (reference standard). This is the diagnostic technique that defines the presence of the condition with the highest known certainty[9,19]. On the other hand, in routine clinical practice, health professionals are confronted with patients who consult them with the result of a test they have already undergone. The probability of being ill from the test results is known as predictive value. This topic will be developed below.

Positive and negative predictive values

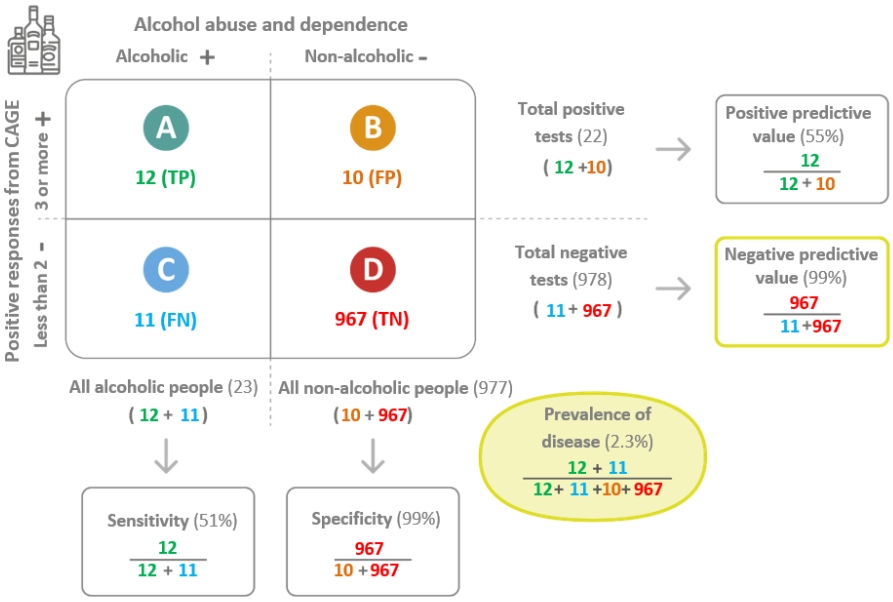
A diagnostic test carries a certain probability that the result correctly categorizes the presence or absence of a disease; this corresponds to predictive values[20]. The positive predictive value is the probability that the diagnostic test correctly identifies sick individuals when it delivers a positive result. In turn, the negative predictive value is the probability that the diagnostic test correctly identifies healthy individuals when it delivers a negative result[21]. Ratios are used to calculate them (Figure 3).

Predictive values are conditioned by the a priori probability of the condition under study[18]. When the a priori probability is low, negative predictive values will be high, and positive predictive values will be low. In this scenario, a negative result of a diagnostic test with a high negative predictive value gives a higher probability to correctly rule out the patient's condition than a positive result to confirm it. On the other hand, when the a priori probability is high, the positive predictive values will be high and the negative predictive values will be low. A positive diagnostic test result with a high positive predictive value gives a higher probability to confirm the condition than a negative result to rule it out[2,16] (Examples 9A and 9B).

Example 9-A.

Suppose one wishes to assess alcohol abuse or dependence in a population with the CAGE questionnaire (Cut-down, Annoyed, Guilty, Eye-opener, whose sensitivity of 51% and specificity of 99% has been estimated previously in validation studies). Locality "A" is a gated community whose community values include abstinence from alcohol. While it cannot be stated that no one drinks alcohol, the estimated prevalence of abuse is low (23/1000) or (2.3%). If we turn our attention to the negative predictive value, it is high (99%) because of the low prevalence of the disease. The effect of prevalence on the negative predictive value in this scenario is indicated by the low number of false negatives in relation to the total number of negative tests. In contrast, the positive predictive value is low (55%). This indicates that it is difficult to confirm a diagnosis with a single test in a low prevalence setting. In this scenario, the effect of prevalence on the positive predictive value is indicated by the high number of false positives in relation to the total number of positive tests (Figure 4).

Figure 4. Contingency table for locality "A".

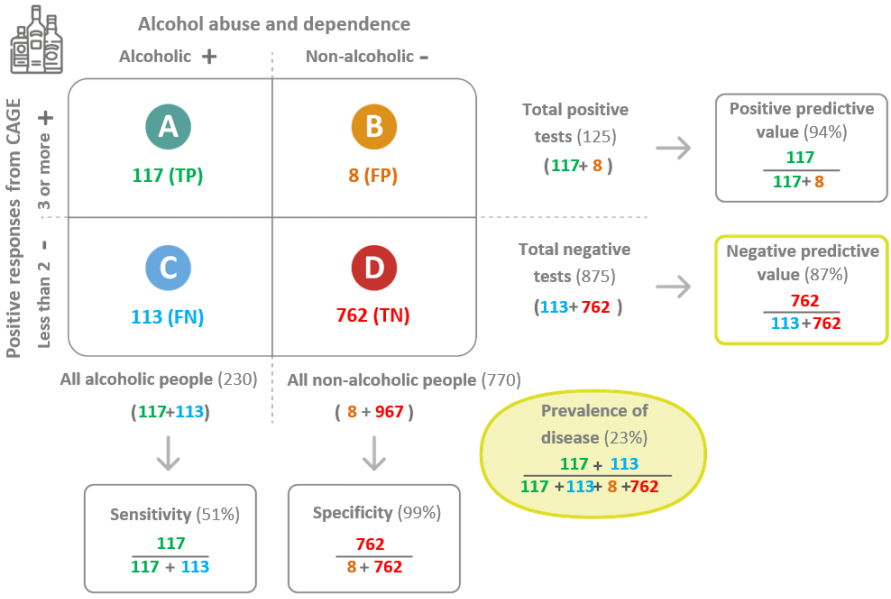


CAGE: Cut-down, Annoyed, Guilty, Eye-opener.
TP: true positives.
FP: false positives.
FN: false negatives.
TN: true negatives.
Source: Prepared by the authors.

Example 9-B.

The same example but in locality "B", the CAGE test has the same sensitivity and specificity values as these are specific to the test used. However, locality "B" has higher alcohol consumption since it is one of the main economic activities (they produce beer), with an estimated abuse or dependence prevalence of 23%. In this context, we can see that the negative predictive value is lower (87%) because it would be more difficult to rule out a diagnosis in a context of high prevalence. The effect of prevalence on the negative predictive value in this scenario is indicated by the high number of false negatives in relation to the total number of negative tests. In contrast, the positive predictive value is higher (94%) because of the high prevalence of the disease. The effect of prevalence on the positive predictive value in this scenario is indicated by the low number of false positives relative to the total number of positive tests. For this reason, a positive result in a context of high prevalence makes the diagnosis more likely compared to a positive result in a context of low disease prevalence[22] (Figure 5).

Figure 5. Contingency table for location "B".



CAGE: Cut-down, Annoyed, Guilty, Eye-opener.
TP: true positives.
FP: false positives.
FN: false negatives.
TN: true negatives.
Source: Prepared by the authors.

Predictive values determine the post-test probability based on the diagnostic test result. However, predictive values are only comparable in populations with a similar prevalence or similar pre-test probability of the condition under study[19].

Likelihood ratios

Likelihood ratios compare the probability of finding a given result (positive or negative) of a diagnostic test in sick individuals in relation to the probability of finding that same result in healthy individuals[16]. Odds ratios are calculated using the sensitivity and specificity of a diagnostic test (Figure 3). Likelihood ratios allow calculation of the probability of disease following the application of a test, adjusting for the different prior probabilities of being ill in different populations[23].

The positive likelihood ratio determines how much more likely it is that the test result will be positive in a sick patient than in a healthy one. In contrast, the negative likelihood ratio determines how much more likely it is that the test result will be negative in a sick patient relative to a healthy one. To facilitate the interpretation of the negative likelihood ratio, the reciprocal of the value calculated for this indicator is used, determining how much more likely it is that the test result will be negative in a healthy patient than in a sick patient (Example 10).

Example 10.

Using the population data for locations A and B from Example 9, we can calculate the CAGE questionnaire's positive and negative likelihood ratio.

Positive likelihood ratio = 0.51/(1 - 0.99) = 51

Negative likelihood ratio = (1 - 0.51)/0.99 = 0.49

The positive likelihood ratio is 51, which means that a sick patient is 51 times more likely to have a positive CAGE questionnaire for alcoholism compared to a healthy patient. The negative likelihood ratio for locations "A" and "B" is 0.49 (to calculate its reciprocal: 1/0.49 ≈ 2), meaning that a healthy patient is two times or twice as likely to have a negative CAGE questionnaire for alcoholism compared to a sick patient.

Positive likelihood ratios can have values between one and infinity and negative likelihood ratios between zero and one. A likelihood ratio of one indicates null utility for discriminating the presence or absence of a condition[23-25] (Table 1).

Table 1. Diagnostic potency.

Capacity	Positive likelihood ratio	Negative likelihood ratio
High	> 10	< 0.1
Moderate	5 to 10	0.1 to 0.2
Low	2 to 5	0.2 to 0.5
Very Low	1 to 2	0.5 to 1

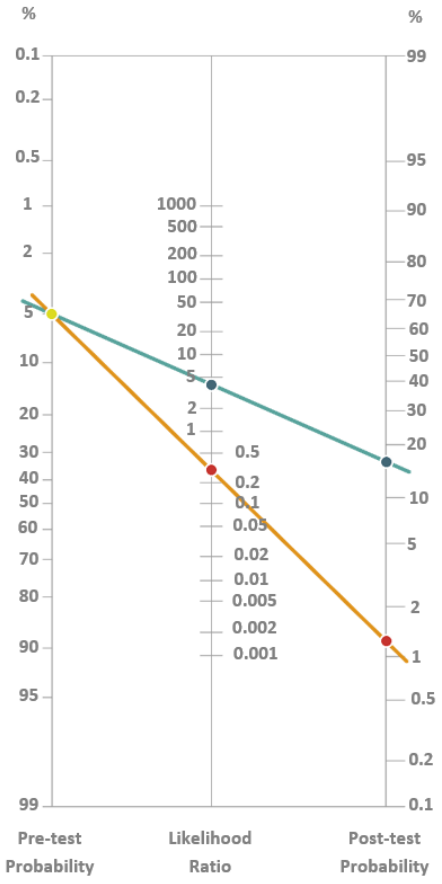
Diagnostic ability of a test according to the likelihood ratio value.
The first column presents the ability or significance of the positive or negative likelihood ratio to modify the pre-test to post-test probability according to the magnitude of its value. Likelihood ratios greater than 10 or less than 0.1 generate huge changes from pre-test to post-test probability. These ratios will be sufficient in most cases to confirm (above the therapeutic threshold) or rule out the condition under study (below the diagnostic threshold)[26]. Odds ratios of 5 to 10 and 0.1 to 0.2 generate moderate changes from pre-test to post-test probability. Odds ratios of 2 to 5 and 0.5 to 0.2 generate small changes in probability.
Source: Prepared by the authors.

The most practical and straightforward way to interpret the likelihood ratios is by applying Bayes' theorem with Fagan's nomogram[27,28]. In this graph, the left column represents the pre-test probability, the middle column the likelihood ratio of the diagnostic test applied and the right column the post-test probability[19]. By extending a straight line joining the values obtained from the first column with that of the second column, it is possible to obtain the result of the third column, corresponding to the probability of having the condition, by means of the diagnostic test result (Example 11).

Example 11.

An 85-year-old female patient consults for morning joint pain in both hands lasting more than one hour. Suspecting rheumatoid arthritis, the physician orders a serological test known as rheumatoid factor. Once the diagnostic test result is available, the Fagan nomogram is used to determine the probability of the disease (Figure 6).

Figure 6. Fagan nomogram of rheumatoid factor.



In this example, the pre-test probability corresponds to the worldwide prevalence in patients over 65 years of age for rheumatoid arthritis, which is approximately 5% (yellow dot in the first column)[29]. The positive likelihood ratio is 4.86 (blue dot in the second column), and the negative likelihood ratio is 0.38 (red dot in the second column)[30]. The post-test probability for the positive outcome is approximately 20% and is obtained by drawing a straight line (light blue line) between the pre-test probability and the positive likelihood ratio. The post-test probability for the negative result is approximately 1%, following the same method (orange line).
Source: Prepared by the authors.

A positive result for the rheumatoid factor, without other signs or symptoms supporting the presence of rheumatoid arthritis, is not sufficient to make the diagnosis, much less to justify treatment[31].

Receiver operating characteristic curve

Some diagnostic tests report their results in continuous or ordinal data, such as blood pressure or glycemia. The cut-off point where the highest sensitivity and specificity exists must be determined when using this type of data, i.e., the place on the curve where the sick are best discriminated from the healthy individuals[32]. However, no exact value separates the sick from the healthy, with overlapping values between the two groups.

Receiver operating characteristic curves are a graphical representation that relates the proportion of true positives (sensitivity) to the proportion of false positives (1 minus specificity) for different possible values of a diagnostic test to determine which value best discriminates between sick and healthy individuals. The receiver operating characteristic curve is constructed from a scatter plot, whose ordinate (y) and abscissa (x) axes correspond respectively to the sensitivity and the complement of the specificity for the different possible outcomes of the diagnostic test. A dotted line is drawn from the lower left corner and the upper right corner of the graph and is called the "reference diagonal" or "non-discrimination line". This reference diagonal corresponds to the theoretical representation of a diagnostic test that does not discriminate between sick and healthy individuals (identical distribution of results for both groups).

The cut-off point that discriminates best between sick and healthy within the receiver operating characteristic curve is the one that achieves the highest sensitivity and specificity at the same time. Graphically, it corresponds to the point closest to the upper left corner of the graph, calculated using the Youden index (sensitivity + specificity - 1)[33]. However, depending on the clinical objective of the diagnostic test, the cut-off point may be different to favor sensitivity or specificity (Example 12).

Example 12.

In the following example, taken and modified from the Clinical Epidemiology book by Feinstein[34], the ergometric test was performed on a sample of two groups of patients, one with proven coronary artery disease and the other without the disease. At the end of the test, the ST-segment elevation was measured (Table 2).

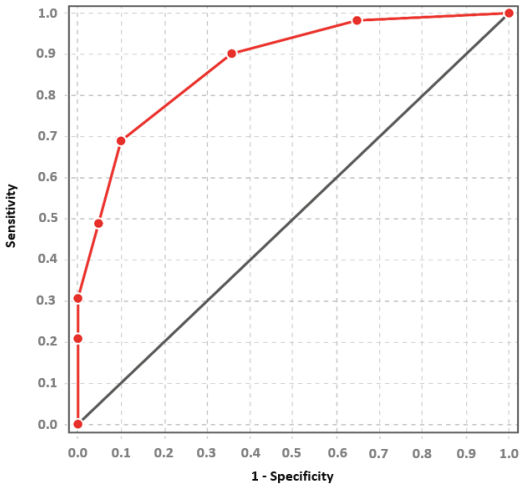
Table 2. Values obtained in the ergometric test.

ST level difference	With coronary artery disease	Accumulated proportion	Without coronary artery disease	Accumulated proportion
≥ 3 millimeters	31	0.21	0	0.00
2.5 - < 3.0	15	0.31	0	0.00
2.0 - < 2.5	27	0.49	7	0.05
1.5 - < 2.0	30	0.69	8	0.10
1.0 - < 1.5	32	0.90	39	0.36
0.5 - < 1.0	12	0.98	43	0.65
< 0.5	3	1.00	53	1.00
Total	150		150	

The first column indicates the ST-segment elevation after the ergometric test was performed. The second column indicates the number of correctly classified patients as sick in the corresponding ST-segment elevation range. The third column reports the cumulative relative frequency of patients correctly classified as sick, i.e., the sensitivity for these ST-segment elevation ranges. The fourth column indicates the number of incorrectly classified patients as sick in the corresponding ST-segment elevation range. Finally, the fifth column shows the cumulative relative frequency of patients incorrectly classified as sick, i.e., the complement of the specificity for these ST-segment elevation ranges.
Source: adapted and modified from Feinstein et al.[34].

The cut-off point that best discriminates between healthy patients and patients with coronary artery disease for this diagnostic test would be the ST-segment elevation greater than or equal to 1.5 millimeters, which has a sensitivity of 0.69 and a specificity of 0.90. However, in clinical practice, the cut-off point used for coronary artery disease is ST-segment elevation greater than 1 millimeter, which has a sensitivity of 0.90 and a specificity of 0.64. This cut-off point privileges sensitivity at the expense of specificity[35,36], since failure to diagnose coronary artery disease (false negative) can be harmful and even fatal for patients. The data obtained in this example are illustrated in a receiver operating characteristic curve (Figure 7).

Figure 7. Operating characteristic curve of the ergometric test receiver.



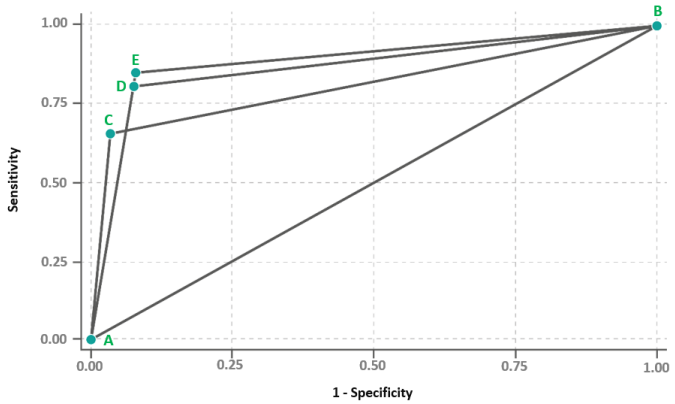
The receiver operating characteristic curve of the ergometric diagnostic test consists of two axes varying from 0 to 1 (0 to 100%), the sensitivity on the vertical axis (y) and the complement of the specificity on the horizontal axis (x). Each red point on the receiver operating characteristic curve are possible cut-off points. The diagonal line in black is called the reference diagonal or non-discrimination line.
Source: adapted and modified from Feinstein et al.[35].

The area under the receiver operating characteristic curve is the global indicator of the accuracy of a diagnostic test, the calculation of which is beyond the scope of this study. This area ranges from 0.5 to 1. At 1, diagnostic tests achieve 100% sensitivity and specificity. An area close to 0.5 means that the diagnostic test cannot discriminate between sick and healthy patients. The area under the receiver operating characteristic curve allows comparisons between two or more diagnostic tests[37], choosing, in general terms, the one with the largest area as the one that best discriminates between sick and healthy patients (Example 12).

Example 12.

Peonim et al.[38] determined that the combined performance of prostate-specific antigen and acid phosphatase is the most accurate diagnostic method for detecting semen in human vaginal specimens. This conclusion was based on comparing both tests' receiver operating characteristic curves, performed separately and together (Figure 8).

Figure 8. Comparison of receiver operating characteristic curves.



This graph compares the receiver operating characteristic curves for acid phosphatase receptor alone (ABC curve), prostate-specific antigen alone (ABD curve) and both tests together (ABE curve). The areas under the receiver operating characteristic curves are respectively 0.8091, 0.8639 and 0.8823. Therefore, the combination of both diagnostic tests has the highest diagnostic accuracy.
Source: adaptation and modification of Figure 1 from Peonim et al. (2013)[38].

Conclusions

Diagnostic tests assist clinical decision-making, and for their analysis, it is essential to understand their properties (sensitivity, specificity, predictive values, and likelihood ratios).

Following Bayes' theorem, from the baseline probability of the individual (pre-test probability) and the properties of the test and its results, we can achieve a new probability for the condition under study.

Receiver operating characteristic curves are valuable tools for evaluating diagnostic tests with non-dichotomous quantitative results, allowing discrimination between two health states.

The correct interpretation of test results can avoid decision-making errors and negative consequences for those subjected to these tests.

Notes

Contributor roles
All authors contributed to the planning and writing of the original manuscript and the introduction, conceptualization, examples, and conclusions of the article.

Competing interests
The authors completed the ICMJE conflict of interest statement and declared that they received no external funds to complete this article; have no financial relationships with organizations that may have an interest in the published article in recent years; and have no other relationships or activities that may influence the publication of the article.

Funding
The authors declare no external funding.

Ethics
This study did not require evaluation by an ethics committee because it worked on secondary sources.

Language of submission
Spanish.

References

1. McGee S. Evidence-Based Physical Diagnosis. 4° Ed. Elsevier; 2017.
2. Molina Arias M. Characteristics of diagnostic tests. Rev Pediatr Aten Primaria. 2013;15(58):169–73. [On line]. | Link |
3. Cobo E, Muñoz P, González JA. Bioestadística para no estadísticos. 1° Ed. Elsevier Masson; 2007.
4. Gross RD. Making Medical Decisions: An Approach to Clinical Decision Making for Practicing Physicians. 1° Ed. ACP Press; 1999.
5. McDonald CJ. Medical heuristics: the silent adjudicators of clinical practice. Ann Intern Med. 1996 Jan 1;124(1 Pt 1):56-62. | CrossRef | PubMed |
6. Pérez Echeverría MP. Psicología del razonamiento probabilístico. 1° Ed. UAM; 1990.
7. Elstein AS. Thinking about diagnostic thinking: a 30-year perspective. Adv Health Sci Educ Theory Pract. 2009 Sep;14 Suppl 1:7-18. | CrossRef | PubMed |
8. Araujo Alonso M. Critical analysis of studies of diagnostic tests: I. Medwave. 2012 Aug 1;12(07):e5465–e5465.
9. Bravo-Grau S, Cruz Q JP. Estudios de exactitud diagnóstica: Herramientas para su Interpretación. Rev chil radiol. 2015;21(4):158–64. [On line]. | Link |
10. Steurer J, Fischer JE, Bachmann LM, Koller M, ter Riet G. Communicating accuracy of tests to general practitioners: a controlled study. BMJ. 2002 Apr 6;324(7341):824-6. | CrossRef | PubMed |
11. Froehling DA, Elkin PL, Swensen SJ, Heit JA, Pankratz VS, Ryu JH. Sensitivity and specificity of the semiquantitative latex agglutination D-dimer assay for the diagnosis of acute pulmonary embolism as defined by computed tomographic angiography. Mayo Clin Proc. 2004 Feb;79(2):164-8. | CrossRef | PubMed |
12. National Lung Screening Trial Research Team, Aberle DR, Adams AM, Berg CD, Black WC, Clapp JD, et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med. 2011 Aug 4;365(5):395-409. | CrossRef | PubMed |
13. Carlos R, Gareen I, Gatsonis C, Gorelick J, Kessler L, Lau J, et al. Standards in the Design, Conduct and Evaluation of Diagnostic Testing for Use in Patient Centered Outcomes Research. PCORI. 2012.[On line]. | Link |
14. Araujo M. Estudios sobre el diagnóstico de las enfermedades. Medwave. 2011 Jul 1;11(07). [On line]. | Link |
15. Burgos D ME, Manterola D C. Assessment of diagnostic test studies. Rev Chil Cir. 2010;62(3):301–8. [On line]. | Link |
16. Mark DB, Wong JB. Decision-making in clinical medicine. In: Harrison's Principles of Internal Medicine. 18° Ed. McGraw Hill Professional; 2011:19-29.
17. Vásquez-Morón JM, Argüelles-Arias F, Pallarés-Manrique H, Ramos-Lora M. Utility of fecal calprotectin in inflammatory bowel disease. RAPD. 2017;40(2). [On line]. | Link |
18. Escrig-Sos J, Martínez-Ramos D, Miralles-Tena JM. Pruebas diagnósticas: nociones básicas para su correcta interpretación y uso. Cirugía Española. 2006 May;79(5):267–73. | CrossRef |
19. Talavera JO, Wacher-Rodarte NH, Rivas-Ruiz R. Investigación clínica II. Estudios de proceso (prueba diagnóstica) [Clinical research II. Studying the process (the diagnosis test)]. Rev Med Inst Mex Seguro Soc. 2011 Mar-Apr;49(2):163-70. | PubMed |
20. Altman DG, Bland JM. Diagnostic tests 2: Predictive values. BMJ. 1994 Jul 9;309(6947):102. | CrossRef | PubMed |
21. Trevethan R. Sensitivity, Specificity, and Predictive Values: Foundations, Pliabilities, and Pitfalls in Research and Practice. Front Public Health. 2017 Nov 20;5:307. | CrossRef | PubMed |
22. Simel D, Rennie D. The Rational Clinical Examination: Evidence-Based Clinical Diagnosis. Mcgraw-Hill; 2008.
23. Deeks JJ, Altman DG. Diagnostic tests 4: likelihood ratios. BMJ. 2004 Jul 17;329(7458):168-9. | CrossRef | PubMed |
24. Pauker SG, Kassirer JP. The threshold approach to clinical decision making. N Engl J Med. 1980 May 15;302(20):1109-17. | CrossRef | PubMed |
25. McGee S. Simplifying likelihood ratios. J Gen Intern Med. 2002 Aug;17(8):646-9. | CrossRef | PubMed |
26. Capurro D, Rada G. The diagnostic process. Rev Med Chil. 2007 Apr;135(4):534–8. [On line]. | Link |
27. Fagan TJ. Letter: Nomogram for Bayes theorem. N Engl J Med. 1975 Jul 31;293(5):257. | CrossRef | PubMed |
28. Aznar-Oroval E, Mancheño-Alvaro A, García-Lozano T, Sánchez-Yepes M. Razón de verosimilitud y nomograma de Fagan: 2 instrumentos básicos para un uso racional de las pruebas del laboratorio clínico [Likelihood ratio and Fagan's nomogram: 2 basic tools for the rational use of clinical laboratory tests]. Rev Calid Asist. 2013 Nov-Dec;28(6):390-1. | CrossRef | PubMed |
29. Artritis reumatoide. Universidad Católica de Chile. [On line]. | Link |
30. Nishimura K, Sugiyama D, Kogata Y, Tsuji G, Nakazawa T, Kawano S, et al. Meta-analysis: diagnostic accuracy of anti-cyclic citrullinated peptide antibody and rheumatoid factor for rheumatoid arthritis. Ann Intern Med. 2007 Jun 5;146(11):797-808. | CrossRef | PubMed |
31. Nicoll D, Lu CM, McPhee SJ. Guide to Diagnostic Tests. 7° Ed. McGraw-Hill Education; 2017.
32. Akobeng AK. Understanding diagnostic tests 3: Receiver operating characteristic curves. Acta Paediatr. 2007 May;96(5):644-7. | CrossRef | PubMed |
33. Böhning D, Böhning W, Holling H. Revisiting Youden's index as a useful measure of the misclassification error in meta-analysis of diagnostic studies. Stat Methods Med Res. 2008 Dec;17(6):543-54. | CrossRef | PubMed |
34. Feinstein AR. Clinical Epidemiology: The Architecture of Clinical Research. 2° Ed. W.B. Saunders Company; 1985.
35. Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clin Chem. 1993 Apr;39(4):561-77. | PubMed |
36. Altman DG, Bland JM. Diagnostic tests 3: receiver operating characteristic plots. BMJ. 1994 Jul 16;309(6948):188. | CrossRef | PubMed |
37. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988 Sep;44(3):837-45. | PubMed |
38. Peonim V, Worasuwannarak W, Sujirachato K, Teerakamchai S, Srisont S, Udnoon J, et al. Comparison between prostate specific antigen and acid phosphatase for detection of semen in vaginal swabs from raped women. J Forensic Leg Med. 2013 Aug;20(6):578-81. | CrossRef | PubMed |

Correspondence to
Angamos 655, Reñaca,
Viña del Mar