sciendo

The Journal of Critical Care Medicine 2020;7(3):241-248

SPOTLIGHT ON...

DOI: 10.2478/jccm-2021-0022

Accuracy of Diagnostic Tests

Ario Santini[1,2*], Adrian Man[1], Septimiu Voidăzan[1]

[1] George Emil Palade University of Medicine, Pharmacy, Science, and Technology of Targu Mures, Romania

[2] Hon. Fellow, The University of Edinburgh

Abstract

Following the outbreak of the coronavirus disease 2019 (COVID-19) pandemic, design, development, validation, verification and implementation of diagnostic tests were actively addressed by a large number of diagnostic test manufacturers. This paper deals with the biases and sources of variation which influence the accuracy of diagnostic tests, including calculating and interpreting test characteristics, defining what is meant by test accuracy, understanding the basic study design for evaluating test accuracy, understanding the meaning of Sensitivity, Specificity, Positive Predictive Value and Negative Predictive Value, and evaluating them numerically, and the ROC curve (or Receiver Operating Characteristic) and the Area under the Curve (AUC).

Keywords: 2019 (COVID-19) pandemic, diagnostic tests, ROC curve, Receiver Operating Characteristic

Received: 11 March 2021 / Accepted: 27 June 2021

Introduction

Following the outbreak of the coronavirus disease 2019 (COVID-19) pandemic, design, development, validation, verification and implementation of diagnostic tests were actively addressed by a large number of diagnostic test manufacturers. No test is ideal and none are 100 per cent reliable. Diagnostic tests establish the presence or absence of disease in order to make treatment decisions. A diagnostic test is carried out on symptomatic individuals or after a screen-positive confirmatory test has been obtained [1].

A new medical test must first undergo a series of be assessments before it can be introduced into general clinical use.

Is it effective? Does the test work in the laboratory? Is it clinically efficient? Does the test work in the patient population of interest? Will the test bring about health outcomes benefits [2]?

Diagnostic accuracy studies

Evaluation of a new test's diagnostic accuracy is carried out to assess how well it discriminates between patients with or without the target disease.

The accuracy of an index test cannot be evaluated without a reference standard. At the commencement of a study, there should be a consensus that the reference standard to be used is more accurate than the index test. More than one acceptable reference standard would be appropriate for use in a test accuracy study.

The test accuracy is defined as a comparison between the disease conditions (Target condition) estimated by a test of interest (Index test) and the best estimate of the actual disease state (Reference standard). It is an unequivocal acknowledgement that most tests make errors even if correctly performed.

A degree of pragmatism may be required when choosing an acceptable reference standard. The most accurate reference standard may not be feasible or ethical. Less accurate methods may have to be used. The reference standard may not always be a gold standard (vide infra); the use of a non-gold or imperfect standard may occur when there is no generally accepted reference standard for the target condition. However, using an imperfect reference standard produces reference standard bias [3].

Method of evaluating the diagnostic accuracy of a medical test with binary test results and dichotomised disease status.

All patients included in the Study-Sample
↓
Index test (New test)
↓
Reference test (Gold Standard)

All patients take the Index test (New test) and the Reference test (Gold Standard) simultaneously or within a short interval to avoid changes in the disease status.

* Correspondence to: Ario Santini, George Emil Palade University of Medicine, Pharmacy, Science, and Technology of Targu Mures, Romania. E-mail: ariosantini@hotmail.com

The Index test versus the Reference test (Gold standard)

	Reference test (Gold standard)	
Index test	Positive	Negative
Positive	True Positive	False positive
Negative	False Negative	True Negative

Depending on the test's resultant characteristics, including sensitivity and specificity standards, one may determine the role the new test can play in the diagnostic schema. It may be deemed better than any existing test, a possible replacement test or used as a triage test.

The basic measures of the diagnostic accuracy of a test are sensitivity and specificity. Other measures are predictive values, likelihood values, overall accuracy, receiver operating characteristic (ROC) curve, area under the ROC curve (AUROC) ROC surface, and volume under the ROC surface (VUS). (vide infra)

Diagnostic Test Characteristics

- Sensitivity (true-positive rate) The proportion of subjects who have the disorder (by the gold standard) who have a positive result by the new test. Specificity (true-negative rate) is the proportion of subjects who do not have the disorder and give a negative test.
- The positive predictive value (PPV) is the proportion of subjects who give a positive test result and have the disease.
- The negative predictive value (NPV) is the proportion of subjects who give a negative test result and do not have the disease.
- The likelihood ratio for a positive test result (LR+) is how much more likely is a positive result found in a person with, as opposed to without, the disease?
- The likelihood ratio for a negative test result (LR-) is how much more likely is a negative result to be found in a person with the disease than not having the disease.
- Accuracy of a test: This is the proportion of subjects who give the correct result.
- A false positive is an error resulting from the incorrect indication of a disease's presence, i.e. the result is positive when, in reality, the patient is disease-free.
- A false negative is an error resulting from the incorrect indication that the patient does not have the disease, i.e. the result is negative when, in reality, the patient has the disease.

Information regarding test accuracy is useful in indicating screening, diagnosis, predisposition, monitoring, prognosis, and drug effectiveness.

- Screening: Which patients have an asymptomatic disease?
- Diagnosis: Which patients have a symptomatic disease?
- Predisposition: Which patients could develop the disease?
- Monitoring:
- Is the disease controlled?
- How advanced is the disease?
- Has the disease recurred?
- Prognosis: Will the disease progress over time?
- Is a drug effective?

The Gold Standard

Comparing the index test results with a reference standard for diagnosing the same target condition in the same participants allows quantifying the above-listed measures.

The **reference standard** could be a **gold standard** that refers to an experimental model that has been thoroughly tested and has a reliable method. It is often the method accepted and used as the current best available test. On occasions, a gold standard may not be used because it is expensive or invasive, or patients do not consent to it. The clinicians may decide not to give the gold test to some patients for medical reasons.

Index test, Reference Standard & Target condition

- Index test: the test under evaluation for accuracy
- Reference standard: the best available standard of identifying the target condition against which the index test results will be compared.
- Target condition: the condition under detection

This can be a pathologically defined condition (e.g. fracture) OR a symptom requiring treatment (e.g. high blood pressure)

Test Population

The population of interest must be clearly defined. It would be incorrect to appraise a diagnostic test using a population that does not represent the target population. e.g. using a population derived from a university student population to appraise a test to be used in care home patients. The ideal sample for a test accuracy study is a consecutive or randomly selected series of patients in whom the target condition is suspected, or for screening studies, the target population.

- The Index test: The index test is the NEW test under evaluation.
- The Reference standard: The reference standard is the standard against which the index test is compared. It is usually the best test currently available but may not necessarily be the test used routinely in practice.

The test accuracy is predicated on a one-sided comparison of the index test results and the reference standard. The reference standard is important in validating the test study's accuracy as there is the assumption that it has a 100% accuracy. This assumption is rarely correct and represents a fundamental flaw in the test. Any inconsistency is presumed to result from errors in the index test. Therefore, the selection of the reference standard is critical to the validity of a test accuracy study, and the definition of the diagnostic threshold forms part of that reference standard. In cases where there is no consensus on the best reference test, a composite reference standard, which is considered a better indicator of actual disease status may be used.

Designing a diagnostic accuracy study

The protocol: The protocol details every step of the study. The problem at this stage should be clearly stated.

- **Selection of participants for the target population:** The target population determines the criteria for including participants in the study. The population is important in deciding on an appropriate study-setting
- **Reference standard:** The reference standard should diagnose the same target condition as the index test. The choice of a reference standard (gold or non-gold) determines the methods used when evaluating the index test.
- **Sample size:** An adequate sample size is critical in making inferences from the statistical analysis
- **Selection of accuracy:** A decision should be made at the protocol stage as to which accuracy measures are to be estimated. This decision will be determined by the test's response (binary or continuous).
- **Eliminate possible bias:** Multiple forms of bias may exist. Anticipating how to avoid or minimise bias is essential.
- **Validation of results:** Validation ensures an understanding of the reproducibility, strengths, and limitations of the study.

Expressing test accuracy

The test accuracy compares the disease condition (target condition) estimated by a test of interest (Index test) and the best estimate of the actual disease stated by the Reference standard. It is indisputable that most tests result in errors, even if properly carried out.

The new test characteristics can be computed with values obtained for Sensitivity and Specificity, The Positive predictive value, The Negative predictive value, the Likelihood ratios, Pre-test probability and Odds, Post-test probability and Odds Receiver operating curve [4].

Each of these values should be calculated.

The four possible outcomes of cross-classification are represented in a diagnostic 2x2 contingency table.

Calculating TP, TN, FP, FN values

Patient number: 1 2 3 4 5 6 7 8 9 10 >>>>>>

Reference results: P P P N P N N P P N >>>>>>

Index (new test) results. P N P N P P N P N P >>>>>>>

TP FN TP TN TP FP TN TP FN FP >>>>>>

Number of TP = 4

Number of FP = 2

Number of TN = 2

Number of FN = 2

Total TP+FP+TN+FN = 10

Note: See text. The reference test is always considered to be 100% though it may not be in reality.

It is against the Reference test results that the Index test results are compared.

Results of diagnostic tests

		Reference standard	
		Positive	**Negative**
Index test	**Positive**	TP	FP
	Negative	FN	TN

2 X 2 table of the results of diagnostic tests.

A false positive is an error in which a test result incorrectly indicates a disease, i.e. the result is *positive* when there is no disease present.

A false negative is an error in which a test result incorrectly indicates no presence of a disease, i.e. the result is *negative* when the disease is present.

Test characteristic, Explanation, Formula

- Sensitivity: (true –positive rate) The proportion of subjects with the disorder by the reference test who give a positive result by the Index (new) test. TP/ TP+FN
- Specificity: (true-negative rate) The proportion of patients without the disorder and who give a negative test. TN/TN+FP
- Positive prediction value: (PPV) The proportion of patients with a positive test who do have the disease. TP/ TP+FP
- Negative prediction value: (NPV) The proportion of subjects with a negative test who do not have the disease. TN/TN+FN
- The likelihood ratio for a positive test result: (LR+) How much more likely is a positive test to be found in a person with the disease compared to being without the disease: sensitivity/ 1- specificity
- The likelihood ratio for a negative test result: (LR-) How much more likely is a negative test to be found in a person with the disease compared to being without the disease. 1- sensitivity / specificity
- False positive rate: Is an error resulting from the incorrect indication of a disease's presence, i.e. the result is *positive* when, in reality, the patient is disease-free. FP/FP+TN
- False negative rate: Is an error resulting from the incorrect indication that the patient does not have the disease i.e. the result is *negative* when, in reality, the patient has the disease. FN/ TP+FN
- Accuracy of a Test: The proportion of the subjects with the correct result. TP+TN/ TP+FP+FN+TN

Helpful aide-memoirs for Sensitivity Specificity and Predictive values

- **SpPin** - when a highly **sp**ecific test is used, a **p**ositive test result tends to rule **in** the disorder.
- **SnNout** - when a highly sensitive test is used, a **n**egative test result tends to rule **out** the disorder.

An explanation of Positive Predictive Value Negative Predictive Value

- **PPV = Positive Predictive Value:** The proportion of those who test positive with the **INDEX TEST** who have the disease?
- **NPV = Negative Predictive Value:** The proportion of those who test negative with the **INDEX TEST** do not have the disease?

Predictive values depend on the **prevalence** of the disorder.

An increase in the **prevalence**** of a disease in a population will increase the positive predictive value. The negative predictive value will decrease.

The likelihood ratio is often more useful than predictive values and can be calculated from sensitivity and specificity numbers. The likelihood ratio remains constant even when the **prevalence** of the disorder changes. [cf. predictive values].

The likelihood ratio indicates the number of times patients with a disease are likely to have a particular test result than patients without the disease.

** Prevalence is the proportion of a particular population affected by a medical condition or disease at a specific time.

The effect of prevalence on the Positive Predictive Value

Prevalence %	VVP %	Sensitivity	Specificity
0.1	1.8	90	95
1	15.4	90	95
5	48.6	90	95
50	**94.7**	**90**	**95**

The ROC Curve

A ROC curve is a performance measurement for the classification problems at various threshold settings. ROC is a probability curve. The area under the curve characterises the degree or measure of separability (Figure 1).

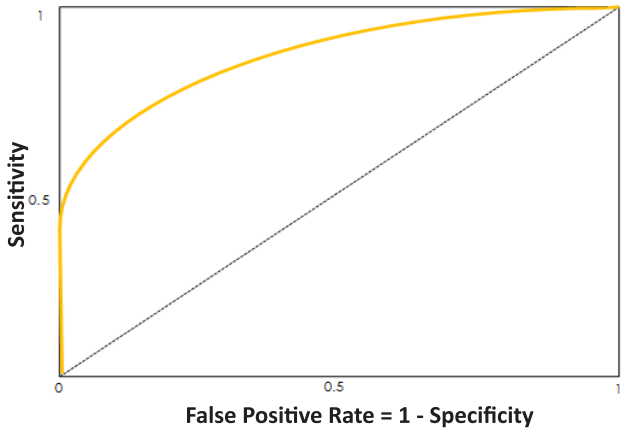

Fig. 1. The ROC curve is composed by calculating the Sensitivity and the False Positive Rate for several thresholds, and plotting them against each other. The False Positive Rate (FPR) or 1 – Specificity is a measurement of how accurate the real negatives are being recorded. The smaller the FPR, the more accurate the identification of the real negative in the data. Sensitivity is recorded on the y axis and is a measure of how accurate people who have a disease are being identified as such.

The Probability Threshold

In medicine, a binary classification problem is knowing the accuracy of a test result patient concerning whether a patient has or has not got a disease. This is a probability question that requires that a threshold is chosen in order to convert this probability into an actual prediction. The threshold should be chosen with care. In medical-related situations, a frequent and important consideration is whether a patient has a disease when he is disease-free. 0.5 probability is the commonly used threshold: when the probability el is greater than 0.5, the prediction is a 1, i.e. the patient has the disease in our case, or , 0 ,the patient does not have the disease.

The probability threshold can be varied depending on the study: this produces different sets of 1s and 0s, and consequently a different set of predictions.

Area under the ROC curve, with standard error and 95% Confidence Interval

This value can be interpreted as follows [5].

- the average value of sensitivity for all possible values of specificity;
- the average value of specificity for all possible values of sensitivity;
- the probability that a randomly selected individual from the positive group has a test result indicating greater suspicion than that for a randomly chosen individual from the negative group.

When the variable under study cannot distinguish between the two groups, i.e. where there is no difference between the two distributions, the area will be equal to 0.5 (the ROC curve will coincide with the diagonal). When there is a perfect separation of the values of the two groups, i.e. there no overlapping of the distributions, the area under the ROC curve equals 1 (the ROC curve will reach the upper left corner of the plot).

The 95% Confidence Interval is the interval in which the true (population) Area under the ROC curve lies with 95% confidence.

P-value

The P-value is the probability that the observed sample Area under the ROC curve is found when in fact, the true (population) Area under the ROC curve is 0.5 (null hypothesis: Area = 0.5). If P is low (P<0.05) then it can be concluded that the Area under the ROC curve is significantly different from 0.5 and that therefore there is evidence that the laboratory test does have an ability to distinguish between the two groups.

		True	False
Predicted labels	Positive	TP	FP
	Negative	FN	TN
		Actual labels	

Defining terms used in a ROC Curve and the AUC

The following metrics that can be extracted from a ROC curve.

The model's **Precision** is calculated using the True row of the Predicted Labels. It is indicative of how good the model is when making a Positive prediction of the

number of patients actually have the disease out of all the Patients that the algorithm predicts are sick.

From the above table, the **Precision** (Positive prediction value) is TP/ TP+FP.

Precision is an important matrix in avoiding mistakes of True predictions, i.e. in the patients who are predicted as having the disease.

Sensitivity (true –positive rate) TP/ TP+FN is calculated using the True Column of the Actual or Real Labels. It indicates how many people who are actually sick are being identified as such. It is a measure of the % of correctly classified True data.

The model's **Specificity** is calculated using the False column of the actual or real labels. It tells how many of the actual healthy patients are being recorded as being without the disease.

Specificity (true-negative rate) TN/TN+FP

It is important in identifying the patients that do not have the disease.

Having defined the metrics that can be used, the probability threshold that gives the best performance is given by using **ROC or Receiver Operating Characteristic** Curve. It represents how Sensitivity and Specificity vary with a change in the probability threshold.

Increasing the sensitivity of a test is generally done to the detriment of specificity and vice versa. It is acknowledged that it is preferable for a screening test for a particular condition to be more sensitive than specific. This means that in fact only a small number of patients go undiagnosed and it is considered acceptable that a certain number of healthy subjects are declared to have that condition.

It encapsulates, in a single, succinct format all of the confusion matrices that would be obtained as the threshold varies from 0 to 1.

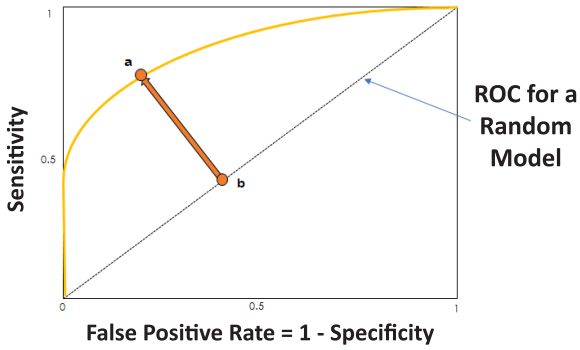

Fig. 2. Representation of the ROC for a random model

The representation of the ROC for a random model

The representation of the ROC for a random model is frequently incorporated in ROC Curves to give a rapid comparison of how well the current model is doing. The further the ROC curve of the data under consideration is distanced from the curve of the random model, the better the distance from point A to point B should be, i.e. Ideally the curve should pass as close as possible to the top-left corner of the diagram. **Figure 2.**

The further the ROC curve of the data under consideration is distanced from the curve of the random model, the better the distance from point A to point B should be.

Explanation of different points of an ROC curve

The points in **Figure 3** explain the meaning of different points of an ROC curve. **Point A** specifies a probability threshold of 1. At this point, the curve produces no True Positives and no False Positives. This means that in-dependently of the probability, every sample

Point A: 0 True Positives & 0 False Positives.

Point B: Some True Positives & 0 False Positives.

Point C: Only True Positives (No false Negatives) & only False Positives (no true Negatives).

Point D: Only True Positives & True Negatives Positives.

Fig. 3. Explanation of different points of an ROC curve

gets classified as False, which is a good threshold to set as the constructed model only makes False predictions. Note that point A is located on the dotted line, which represents a purely random classifier. **Point B** is at a threshold value where some True Positive values are acquired and samples that have a high probability of being positive get correctly classified as such. There are no False Positives at this point. **Point C** is set at the threshold value of 0. Everything is getting labelled as True. (cf point A). If the threshold is set here, the model only creates true predictions. **Point D** is the point of optimal performance; only True Positives and True Negatives are recorded; every prediction is correct. The ideal aim is to get as close as possible to that top left corner, but it is extremely idealistic and unlikely that a ROC curve would reaches this point.

Pragmatically, the aim is to identify a point between B and C in the curve that fulfils success on the 0s and success on the 1s, and picking the threshold related to that point.

The Area under the ROC Curve: (AUC)

The area under the ROC curve (AUC is a measurement from values of 0.5 (random classifier) to 1 (perfect classifier). It signifies how well the model classifies the True and False data points. The greater AUC results in the ROC approaching the desired top-left corner. (vide supra) **Figure 4.**

Conclusion; The more area under **our ROC curve, the better the model is.**

An example for ROC curves of age and ESR in cancer

Figure 5 is an example for ROC curves of age and ESR in cancer. For age the area under the curve is 0.684, and for ESR = 0.690. It can be seen how the curves are closer to the reference line (area = 0.5) than to the upper left corner, the point of maximum accuracy of the test [6].

As most diagnostic tests are far from perfect, often a single test is insufficient. For this reason, clinicians use multiple diagnostic tests, administered either in parallel or in series. In the case of a patient with polyarthritis, for example, it can be said that she has lupus whether she has a malarial rash, or nephrotic syndrome, or thrombocytopenia, or pleural effusion, or antinuclear antibodies (ANA), etc. By applying the tests *in parallel*, the sensitivity is increased, practically, no patients with lupus are lost, but when specificity is decreased, patients diagnosed with lupus may actually have another disease, so the test was false positive. When a battery of tests is applied *in series*, the result is considered positive when all the tests that making up the battery are positive, and negative when at least one is negative. Taking the same example as lupus, this diagnosis is made when the patient with polyarthritis has at the same time malarial rash, nephrotic syndrome, thrombocytopenia, pleural effusion and ANA. We see, therefore, how this method increases the specificity (a patient who meets all these criteria, certainly has lupus), losing, instead, sensitivity (patients who do not have all these manifestations of the disease, but only some among them).

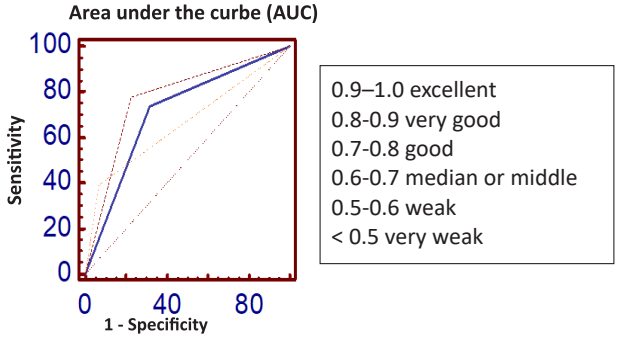

Fig. 4. The area under the ROC curve (AUC) is a measurement from values of 0.5 (random classifier) to 1 (perfect classifier). It signifies how well the model classifies the True and False data points. The greater AUC results in the ROC approaching the desired top-left corner.

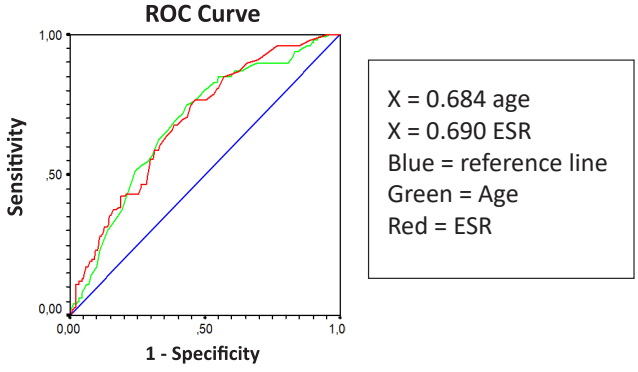

Fig. 5. An example for ROC curves of age and ESR in cancer. For age the area under the curve is 0.684, and for ESR = 0.690. It can be seen how the curves are closer to the reference line (area = 0.5) than to the upper left corner, the point of maximum accuracy of the test.

Summary

The test accuracy is the comparison between the disease state estimated by a test of interest, the Index test, and the best estimate of the true disease state provided by the Reference standard.

Interpretation of numerical test accuracy metrics requires consideration of the number and consequences of test errors.

To decide which dimension of test accuracy is more important in a testing situation, the consequence of being an Index test positive or an Index test negative need to be considered.

Conflict of interest

None to declare.

References

1. STARD statement: STARD initiative (STAndoRDs for the Reporting of Diagnostic accuracy studies) aims to improve the accuracy and completeness of diagnostic accuracy studies. The STARD statement consists of a checklist of 25 items which can be viewed on their web site http:// www.stard-statement.org/
2. Bossuyt PM, Reitsma JB, Linnet K, Moons KG. Beyond diagnostic accuracy: the clinical utility of diagnostic tests. Clinical chemistry. 2012; 58(12):1636–43.
3. Bossuyt PMI L., Craig J, Glasziou P. Comparative accuracy: assessing new tests against existing diagnostic pathways. British Medical Journal. 2006; 332(7549):1089–92.
4. Altman DG, Bland JM. Diagnostic tests 1: Sensitivity and specificity. British Medical Journal. 1994; 308 (6943):1552.
5. Zhou XH., McClish DK, Obuchowski NA. Statistical Methods in Diagnostic Medicine. John Wiley and Sons. 2002: 464 pp.
6. 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 basedon a secondary care university hospital in Romania. Eur J Intern Med. 2006; 17:28-31.