Statistical Assessment for Measuring Diagnostic Accuracy

Sunmeet Matkar

IGMPI New Delhi
E-mail: matkar.sunmeet@gmail.com

Abstract—*Diagnostic accuracy is associated with the potential of a diagnostic test to distinguish between good health and the target condition. This ability to distinguish can be enumerated by the measures of diagnostic accuracy involving sensitivity and specificity, positive and negative predictive values, prevalence, likelihood ratio, the area under the ROC curve, Youden's index and diagnostic odds ratio. Various measures of diagnostic accuracy are associated with the varied aspects of diagnostic procedure, whereas certain measures are applied to evaluate the distinguishing attribute of the diagnostic test, others are used to examine its predictive ability. The measures of diagnostic accuracy are very sensitive to the design of the study. Studies failing to adhere to stringent methodological standards usually over-or under-estimate the indicators of test performance as well as they limits the applicability of the results of the study. The initiative "Standards for Reporting Diagnostic accuracy studies" (STARD) was developed to contribute to the completeness and transparency of reporting of diagnostic accuracy studies. STARD statement should be annexed into the Instructions to authors by scientific journals and authors should be encouraged to use the checklist whenever reporting their studies on diagnostic accuracy. These developments could make a significant difference in the quality of reporting of studies of diagnostic accuracy. They will serve to provide the accurate possible evidence for the state-of-the-art patient care services. The current article summarizes selected fundamental yet indispensible definitions and characteristics of the measures of diagnostic accuracy.*

Keywords: *(Area under Curve (AUC), Diagnostic Odds Ratio (DOR), Diagnostic accuracy, Likelihood ratio, STARD.*

1. INTRODUCTION

Diagnostic accuracy of a diagnostic test provides an answer to this pertinent research question: "How well this test discriminates between two stages of a disease; disease and health?" This ability to distinguish can be calculated by the measures of diagnostic accuracy [1]:

- Sensitivity and Specificity
- Positive and Negative predicative values (PPV, NPV)
- Likelihood ratio
- Area under the ROC curve
- Youden's index
- Diagnostic odds ratio (DOR)

Various measures of diagnostic accuracy are associated with the varied aspects of diagnostic procedure. Some measures are used to assess the discriminative property of the test; others are used to assess its predictive ability. Certain measures are applied to evaluate the distinguishing attribute of the diagnostic test while others are used to examine its predictive ability. Measures of diagnostic accuracy are very sensitive to the characteristics of the population in which the test accuracy is evaluated. Some measures largely depend on the disease prevalence, while others are highly sensitive to the spectrum of the disease in the studied population. It is therefore of utmost importance to know how to interpret them as well as when and under what conditions to use them.

1.1 SENSITIVITY AND SPECIFICITY: An accurate diagnostic procedure has the ability to completely distinguish patients with and without disease. The two basic measures of quantifying the diagnostic accuracy of a test are the sensitivity and specificity. Sensitivity is the ability of a test to detect the disease when it is truly present, whereas specificity is the probability of a test to exclude the disease status in patients who do not have the disease. Thus, sensitivity is given by the ratio of true positives (true positives false negatives), and specificity is given by the ratio of true negatives/ (true negative false positives). In the example given in Table 1, the sensitivity is 90% (540/600) and specificity is 60% (120/200).

Table 1: Test Results by Disease Status with Disease Prevalence of 75%

	Disease status (Gold Standard)		
Test results	**Present**	**Absent**	**Total**
Positive	True Positive (540)	False Positive (80)	620
Negative	False Negative (60)	True Negative (120)	180
Total	600	200	800

In describing a diagnostic test, one needs to report both sensitivity and specificity because they are inherently linked in that as the value of one increases, the value of the other decreases. The values of a perfect test which are above the cut-off are always suggesting the disease, while the values below the cut-off are always eliminating the disease. It is important to note that the values above the cut-off are not always indicative of a disease since subjects without disease can also sometimes have elevated values. Such elevated values of certain parameter of interest are called false positive values (FP). On the other hand, values below the cut-off are mainly found in subjects without disease: false negative values (FN). Therefore, the cut-off divides the population of examined subjects with and without disease in four subgroups considering parameter values of interest: true positive (TP) – subjects with the disease with the value of a parameter of interest above the cut-off, false positive (FP) – subjects without the disease with the value of a parameter of interest above the cut-off, true negative (TN) – subjects without the disease with the value of a parameter of interest below the cut-off and false negative (FN) – subjects with the disease with the value of a parameter of interest below the cut-off. The first step in the calculation of sensitivity and specificity is to make a 2x2 table with groups of subjects divided according to a gold standard or (reference method) in columns, and categories according to test in rows (Table 1). Sensitivity is expressed in percentage and defines the proportion of true positive subjects with the disease in a total group of subjects with the disease (TP/TP+FN). Actually, sensitivity is defined as the probability of getting a positive test result in subjects with the disease. Hence, it relates to the potential of a test to recognize subjects with the disease. Specificity is a measure of diagnostic test accuracy, complementary to sensitivity. It is defined as a proportion of subjects without the disease with negative test result in total of subjects without disease (TN/TN+FP). In other words, specificity represents the probability of a negative test result in a subject without the disease.

1.2 POSITIVE AND NEGATIVE PREDICTIVE VALUES: Positive predictive value (PPV) is the probability that a patient has the disease given that the test results are positive, and the negative predictive value (NPV) is the probability that a patient does not have the disease given that the test results are indeed negative. PPV is therefore given by the ratio of true positives/ (true positives + false positives), and NPV is given by the ratio of true negatives/ (true negatives + false negatives). Unlike sensitivity and specificity, predictive values are largely dependent on disease prevalence in examined population. Therefore, predictive values from one study should not be transferred to some other setting with a different prevalence of the disease in the population. Prevalence affects PPV and NPV differently. PPV is increasing, while NPV decreases with the increase of the prevalence of the disease in a population. Whereas the change in PPV is more substantial, NPV is somewhat weaker influenced by the disease prevalence.

1. 3 LIKELIHOOD RATIO: Likelihood ratio is a very useful measure of diagnostic accuracy. It is defined as the ratio of expected test result in subjects with a certain state/disease to the subjects without the disease [2]. The LR is really the ratio of sensitivity to (100-Specificity). Therefore, it is independent of prevalence of the disease. The magnitude of the LR informs about the certainty of a positive diagnosis. As a general guideline, a value of LR 1 indicates that the test result is equally likely in patients with and without the disease, values of LR> 1 indicate that the test result is more likely in patients with the disease and values of LR < 1 indicate that the test result is more likely in patients without the disease.2 Likelihood ratio for positive test results (LR+) tells us how much more likely the positive test result is to occur in subjects with the disease compared to those without the disease (LR+=(T+│B+)/(T+│B-)). LR+ is usually higher than 1 because is it more likely that the positive test result will occur in subjects with the disease than in subject without the disease. Likelihood ratio for negative test result (LR-) represents the ratio of the probability that a negative result will occur in subjects with the disease to the probability that the same result will occur in subjects without the disease. Therefore, LR- tells us how much less likely the negative test result is to occur in a patient than in a subject without disease. (LR-= (T-│B+)/ (T-│B-)). LR- is usually less than 1 because it is less likely that negative test result occurs in subjects with than in subjects without disease. LR- is a good indicator for ruling-out the diagnosis.

1.4 ROC Curve: There is a pair of diagnostic sensitivity and specificity values for every individual cut-off [3]. To construct a ROC graph, we plot these pairs of values on the graph with the 1-specificity on the x-axis and sensitivity on the y-axis (Fig. 1).

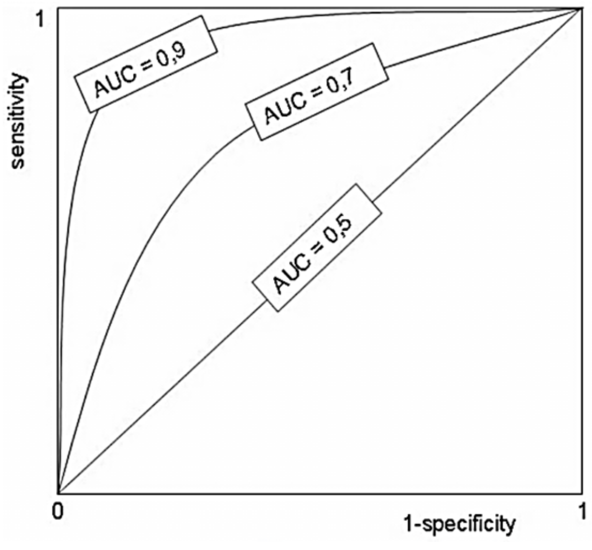

Fig. 1: ROC curve

The shape of a ROC curve and the area under the curve (AUC) helps us estimate how high the discriminative power of a test is. The closer the curve is located to upper-left hand corner and the larger the area under the curve, the better the test is at discriminating between diseased and non-diseased.

The area under the curve can have any value between 0 and 1 and it is a good indicator of the goodness of the test. A perfect diagnostic test has an AUC 1.0. Whereas a nondiscrimination test has an area 0.5. Generally we can say that the relation between AUC and diagnostic accuracy applies as described in Table 2.

Table 2: Relationship between the area under the ROC curve and diagnostic accuracy

Area	Diagnostic accuracy
0.9- 10.	Excellent
0.8 – 0.9	Very good
0.7 – 0.8	Good
0.6 – 0.7	Sufficient
0.5 – 0.6	Bad
< 0.5	Test not useful

AUC is a global measure of diagnostic accuracy. It tells us nothing about individual parameters, such as sensitivity and specificity. Out of two tests with identical or similar AUC, one can have significantly higher sensitivity, whereas the other significantly higher specificity. Furthermore, data on AUC state nothing about predicative vales and about the contribution of the test in ruling-in and ruling-out a diagnosis. Global measures are there for general assessment and for comparison of two or more diagnostic tests. By the comparison of areas under the two ROC curves we can estimate which one of two tests is more suitable for distinguishing health from disease or any other two conditions of interest. It should be pointed that this comparison should not be based on visual nor intuitive evaluation (4). For this purpose we use statistic tests which evaluate the statistical significance of estimated difference between two AUC, with previously defined level of statistical significance (P).

1.5. Diagnostic odds ratio (DOR): Diagnostic odds ratio is also one global measure for diagnostic accuracy, used for general estimation of discriminative power of diagnostic procedures and also for the comparison of diagnostic accuracies between two or more diagnostic tests. DOR of a test is the ratio of the odds of positivity in subjects with disease relative to the odds in subjects without disease (5). It is calculated according to the formula: DOR = (TP/FN) / (FP/TN). DOR depends significantly on the sensitivity and specificity of a test. A test with high specificity and sensitivity with low rate of false positives and false negatives has high DOR. With the same sensitivity of the test, DOR increases with the increase of the test specificity. For example, a test with sensitivity > 90% and specificity of 99% has a DOR greater than 500. DOR does not depend on disease prevalence; however like sensitivity and specificity it depends on criteria used to define disease and its spectrum of pathological conditions of the examined group (disease severity, phase, stage, comorbidity etc.).

1.6. Diagnostic effectiveness (accuracy): It is expressed as a proportion of correctly classified subjects (TP+TN) among all subjects (TP+TN+FP+FN) [4]. It is affected by the disease prevalence. With the same sensitivity and specificity, diagnostic accuracy of a particular test increases as the disease prevalence decreases.

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

STARD initiative was published in 2003. It was a very important step toward the improvement the quality of reporting of studies of diagnostic accuracy [5]. According to some authors, the quality of reporting of diagnostic accuracy studies did not significantly improve after the publication of the STARD statement, whereas some others hold that the overall quality of reporting has at least slightly improved, but there is still some room for potential improvement. Editors of scientific journals are encouraged to include the STARD statement into the Journal Instructions to authors and to oblige their authors to use the checklist when reporting their studies on diagnostic accuracy. This way the quality of reporting could be significantly improved, providing the best possible evidence for health care providers, clinicians and laboratory.

2. SUMMARY

Studies designed to measure the performance of diagnostic tests are important for patient care and health care costs. Attention must be given to include proper representation of patients with the disease or condition of interest along with healthy participants to ensure that the study results are generalizable to the population of interest. Extrapolation of results obtained from one study to other populations requires a good understanding of the underlying prevalence and its impact on the estimates of PPV and NPV.

3. ACKNOWLEDGEMENTS

I acknowledge the supported provided by IGMPI.

REFERENCES

[1] Irwig L, Bossuyt P, Glasziou P, Gatsonis C, Lijmer J, "Designing studies to ensure that estimates of test accuracy are transferable", *BMJ,* 324, 7338, 2002, pp.669-671.

[2] Deeks JJ, Altman DG, "Diagnostic tests 4: likelihood ratios" *BMJ,* 17, 329(7458), 2004, pp.168-169.

[3] Obuchowski NA, Lieber ML, Wians FH Jr, "ROC curves in clinical chemistry: uses, misuses, and possible solutions", *Clin Chem*, 50(7), 2004, pp. 1118-1125.

[4] Bossuyt PM, "Clinical evaluation of medical tests: still a long road to go", *Biochemia Medica*, 16(2)89, 2006, pp. 228.

[5] Bossuyt PM, "The quality of reporting in diagnostic test research: getting better, still not optimal", *Clin Chem*, 50(3), 2004, pp. 465-466.