Review Article

Jale Karakaya*

Evaluation of binary diagnostic tests accuracy for medical researches

[Tıbbi araştırmalar için iki sonuçlu tanı testlerinin doğruluğunun değerlendirilmesi]

https://doi.org/10.1515/tjb-2020-0337
Received July 8, 2020; accepted November 6, 2020;
published online November 30, 2020

Abstract

Objectives: The aim of this study is to introduce the features of diagnostic tests. In addition, it will be demonstrated which performance measures can be used for diagnostic tests with binary results, the properties of these measures and how to interpret them.
Materials and Methods: The evaluation of the diagnostic test performance measures may differ depending on whether the test result is numerical or binary. When the diagnostic test result is continuous numerical data, ROC analysis is often utilized. The performance of a diagnostic test with binary results are usually evaluated using the measures of sensitivity and specificity. However, there are some important measures other than these two measures for binary test results. These measures are predictive values, overall accuracy, diagnostic odds ratio, Youden index, and likelihood ratios.
Results: A hypothetical data has been produced based on the studies conducted on the performance of rapid tests (Specific IgM/IgG) according to the RT-PCR test for Covid 19 in the literature. An example of a diagnostic test (Specific IgM/IgG) with a binary result is given and all measurements and their confidence interval are obtained for this data. The performance of rapid test was examined and interpreted.
Conclusion: It is important to design and evaluate the performance of diagnostic/screening tests for health care. In this review, some basic definitions, performance measures that can be used only in evaluating the diagnostic tests with binary results and their confidence intervals are mentioned. Having many different measures provides different interpretations in the evaluation of test performance. Accurately predicting the performance of a diagnostic test depends on many factors. These factors can be study design, criteria of participant selection, sample size calculation, test methods etc. There are guidelines that ensure that all information regarding the conditions under which the study was conducted is in report, in terms of such factors. Therefore, these guidelines are recommended for use of the checklist by many publishers.

Keywords: binary diagnostic test; diagnostic odds ratio; likelihood ratios; overall accuracy; predictive values; sensitivity; specificity.

Öz

Amaç: Bu çalışmada amaç, tanı testlerinin özelliklerini tanıtmak. Buna ek olarak, İki sonuçlu tanı testleri için hangi performans ölçülerinin kullanabileceği, bu ölçülerin özelikleri ve nasıl yorumlanacağı gösterilecektir.
Gereç ve Yöntem: Tanı testi sonucunun sayısal veya iki sonuçlu olmasına göre test performans ölçülerinin değerlendirilmesi farklılık gösterebilir. Tanı testi sonucu sürekli sayısal olduğunda sıklıkla ROC analizinden yararlanılır. İki sonuçlu bir tanı testinin performansı genellikle duyarlılık ve seçicilik ölçüleri ile değerlendirilir. Ancak, iki durumlu testler için bu iki ölçü dışında başka bazı önemli ölçüler de vardır. Bu ölçüler; kestirim değerleri, genel doğruluk oranı, tanı odds oranı, youden indeksi, olabilirlik oranlarıdır.
Bulgular: Literatürde bulunan Covid 19 hastalığı için RT-PCR testine göre hızlı testin (Specific IgM/IgG) performansı üzerine yapılan çalışmalar temel alınarak hipotetik

***Corresponding author: Jale Karakaya**, Ph.D, Department of Biostatistics, School of Medicine, Hacettepe University, Sihhiye, Ankara, Turkey, E-mail: jalekarakaya@gmail.com. https://orcid.org/0000-0002-7222-7875

bir veri üretilmiştir. İki durumlu bir tanı testine (Specific IgM/IgG) örnek verilmiş ve bu veri için tüm ölçüler ve onların güven aralıkları elde edilmiştir. Hızlı testin performansı incelenmiş ve yorumlanmıştır.

Sonuç: Sağlık hizmetleri için tanı ve tarama testlerinin performanslarının tasarlanması ve değerlendirilmesi önemlidir. Bu derlemede, bazı basit tanımlar, sadece iki sonuçlu tanı testlerinin değerlendirilmesinde kullanılan performans ölçüleri ve bunlara ait güven aralıklarından söz edilmiştir. Birçok farklı ölçü olması performansın değerlendirilmesinde farklı yorumlamalar sağlar. Bir tanı testinin performansının doğru olarak kestirilmesi birçok faktöre bağlıdır. Bu faktörler, çalışma tasarımı, katılımcıların seçim kriterleri, örneklem büyüklüğü hesaplaması, test yöntemleri vb. olabilir. Böyle faktörler açısından, çalışmanın yapıldığı koşullara ilişkin tüm bilginin çalışma raporunda olmasını sağlayan kılavuzlar bulunmaktadır. Bu nedenle, birçok yayınevi tarafından bu kontrol listelerinin kullanılması için bu kılavuzlar önerilmektedir.

Anahtar Sözcükler: duyarlılık; genel doğruluk oranı; İki sonuçlu tanı testi; kestirim değerleri; olabilirlik oranı; seçicilik; tanı odds oranı.

Introduction

Various diagnostic and laboratory tests are used in a medical process of deciding whether a person has a specific disease. The results of diagnostic tests about a person may not always be accurate. In other words, they cannot distinguish patients and healthy peoples 100% accurately. While some tests are perfect, such as reference tests, completely distinguishing the diseased from the non-diseased subjects, others may lead to mis-classifications (wrong diagnosis) due to indefinite outcomes [1, 2]. Diagnostic tests which has 100% accurate results are known as "gold standard" tests. Although outcomes of some tests cannot accurately discriminate the non-diseased and diseased, they are still used as a reference test for being the best available preference. It may not always be possible to use reference tests since they are costly, risky, difficult, and so on. Consequently, imperfect tests with low cost, fast, and low risk are frequently used to make a diagnosis. These tests are also known as index test. Index test is a diagnostic test that is being evaluated against a reference standard test in a study of test accuracy. Then, how reliable are the outcomes of such tests? The answer of this question is related to what extent the test applied yields accurate results. The fact that index tests other than reference tests can be used in the diagnostic process depends on knowing how accurately it can distinguish patients and healthy people. It is quite important in medical practice to know in advance the possible accuracy or inaccuracy of the results of these indefinite tests [1, 2]. Some measures have been developed to assess the performance of these tests in terms of their discrimination accuracy. The measures that can be used in statistical evaluation processes vary with respect to the purpose and whether the outcome of diagnostic test is numerical or categorical (i.e. positive-negative). The accuracy of a diagnostic test when the outcome is quantitative (numerical) or ordinal is usually evaluated by a receiver operating characteristic (ROC) analysis [3, 4]. Although the results of some tests are numerical, they can still be assessed as negative or positive by applying a cut-off value.

This article deals only with performance measures that can be used to assess the classification performance of diagnostic tests whose outcome is binary or numerical results have been transformed into binary. For assessing the success of a diagnostic test in classification, it is necessary to have both reference and index tests applied independently to each individual and to evaluate outcomes [5, 6].

Statistical measures for diagnostic accuracy assessment

Many measures have been developed to assess the accuracy of an index test. Each of these measures has different interpretation and domain of use. It is appropriate to use the measures listed below in cases where the test outcome is binary (positive–negative).

- Sensitivity
- Specificity
- Positive predictive value (PPV)
- Negative predictive value (NPV)
- Overall test accuracy
- Diagnostic odds ratio (DOR)
- Youden Index (YI)
- Positive Likelihood ratio (LR+)
- Negative Likelihood ratio (LR–)

If true disease status and the outcome of imperfect diagnostic test results are binary, basic diagnostic measures used in assessing the performance of diagnostic tests are sensitivity, specificity, false positive rate and false negative rate [7]. To obtain these measures there is need for a 2×2 contingency table relating to individuals to whom both tests are applied as given in Table 1 below.

In this table, the true disease status **(D)** obtained through gold standard test is denoted as D+ in the presence

Table 1: 2 × 2 contingency table for the diagnostic test with binary result.

Diagnostic (index) test result	True disease status gold standard test (or reference test)		Total
	Disease present (D+)	Disease absent (D−)	
Test positive (T+)	True positive (TP)	False positive (FP)	TP + FP
Test negative (T−)	False negative (FN)	True negative (TN)	FN + TN
Total	TP + FN	FP + TN	N (TP + FN + FP + TN)

of disease and D− in the absence of it. Similarly, while the result obtained by the diagnostic test is (**T**), the presence of disease (positive outcome) is denoted by T+ and its absence (negative) by T−.

The probability of a subject randomly selected from N individuals to be with a disease (marginal probability) is defined by the following equation:

$$P(D+) = (TP + FN)/N$$

When this probability is obtained from a study with cross-sectional design it is used in estimating prevalence. The probability of positive result from an index test is obtained by:

$$P(T+) = (TP + FP)/N.$$

From this 2 × 2 table Sensitivity (True Positive Rate), Specificity (True Negative Rate), False Positive Rate and False Negative Rate can be easily obtained as basic performance measures of a test. All these measures represent conditional probability which is denoted as P(A|B) meaning the probability of event A given that the event B is known.

$$P(A \mid B) = \frac{P(A \text{ ve } B)}{P(B)}$$

In performance measures, these two events are considered as index test result (A) and true disease status (B).

Sensitivity and specificity

Sensitivity (P(T+|D+)) is the probability that the index tests yields a positive result (T+) for a diseased subject (D+).

Specificity (P(T−|D−)) is the probability that the index tests yields a negative result (T−) for a healthy subject (D−).

False Positive rate P(T+|D−) is the probability that the index tests yields a positive result (T+) for a healthy subject (D−).

False Negative rate P(T−|D+) is the probability that the index tests yields a negative result for a diseased subject (D+).

$$\text{Sensitivity}\,(TPR) = TP/(TP + FN)$$

$$\text{Specificity}\,(TNR) = TN/(FP + TN)$$

$$\text{False Positive Rate}\,(FPR) = FP/(FP + TN) = 1 - \text{Specificity}$$

$$\text{False Negative Rate}\,(FNR) = FN/(TP + FN) = 1 - \text{Sensitivity}$$

In diagnostic tests, sensitivity and specificity values close to 1 indicate higher discriminative power. The test is considered as perfect when these values are both equal to 1. However, this is relevant to gold standard tests only. In most tests values are under 1 and such tests are known as “imperfect”. There may be tests of different characteristics in diagnosing the same disease and these tests may have their different performance measures. For example, a test with higher sensitivity value relative to another may have lower value in specificity. High sensitivity value for a test means lower FNR (=1 − sensitivity) value for the same test. In this case, it shows that the test is not much likely to inaccurately diagnose actually diseased subject as healthy and would not miss actually diseased. Hence, negative results of this test will be more reliable. When specificity value is high, FPR will be low. Then, positive outcomes of an index test with high specificity value will be more reliable [2].

Tests diagnosing the same disease and displaying such differing performances may be preferred depending on the purpose for which they are used. While some tests are used for diagnosis, others are used for screening. Although diagnostic and screening tests are used for different purposes, the same mathematical process are used to assess the accuracy of these test.

Screening tests can be rapidly applied in a given community and used in revealing diseases that were not known earlier. The objective is to identify and diagnose suspected cases correctly as early as possible. These tests in general do not claim to reach definitive diagnosis and they are needed to identify a disease at its early stage and to start early treatment. In screening tests, it is critical not to miss diseased individuals as far as possible. Hence, tests that are used for screening purposes must have very low

FNR, and thus quite high sensitivity value. FNR values in these tests may be high; but often another confirmation test is applied to subject with positive result in the first test. Confirmation tests may be thought of as having higher specificity or a reference test. It is on the basis of this test that false positive subjects are distinguished from diseased subject at this stage.

In tests used for the purpose of diagnosis it is desired to have as high specificity value as possible. High specificity means low FPR. Generally, individuals with positive diagnosis from this test need to undergo a further process. If these subjects are associated with false positive diagnosis they may have to undergo more advanced examinations, receive unnecessary treatment or operation although they are in fact healthy. It is necessary to use a test with a high sensitivity value in order to prevent subjects who have positive diagnosis despite being healthy from being exposed to these procedures unnecessarily.

Both sensitivity and specificity are among measures that are not affected by prevalence. However, these measures may be varying depending on the disease spectrum [8–11]. In other words, there may be some factors affecting the outcome of the diagnostic test (i.e. sex, age, body mass index (BMI), etc.). For example, a diagnostic test may not display the same sensitivity or specificity values in males and females. An example is the study conducted by Karakaya et al. [11]. This study found "waist to hip ratio" and "BMI" as factors affecting Fasting Plasma Glucose's sensitivity value in diagnose of diabetes, and "age", "hyperlipidaemia" and "family history" as factors affecting specificity, and then examined the performance of the test in sub-groups.

In test performance studies, there is a need for identifying factors affecting sensitivity and specificity values, and sub-group analyses should be performed according to these factors since such analyses provide much further information about the test [11].

Overall accuracy (OA)

The overall accuracy rate is obtained by dividing concordance cells of both tests (true positive + true negative) by total number of subjects involved.

$$\text{Overall accuracy} = (TP + TN)/(TP + FP + FN + TN)$$

The overall accuracy does not allow for examining the performance of negative or positive results of the test; its focus is on the performance of the test with respect to the accuracy of classification.

As an example, let's suppose we would like to compare performances of two different diagnostic tests applied to 100 diseased and 100 non-diseased subjects in order to diagnose the same disease. While the OA of the test with 60% sensitivity and 80% specificity is 70% ([60 + 80]/200), the OA of another test with 80% sensitivity and 60% specificity will again be 70% ([80 + 60]/200). While two tests have different sensitivity and specificity, taking the value of overall accuracy only may lead to the conclusion that both tests have the same performance. Yet, we know that these tests have different sensitivity and specificity. The OA measure alone therefore misses the chance of assessing their performance in terms of negative and positive cases. It is also a disadvantage that it is affected by the prevalence of the disease.

There are also measures where both sensitivity and specificity are used together. These are positive and negative predictive values, diagnostic odds ratio, Youden index, positive and negative likelihood ratios, each suggesting different interpretations. The details and equivalences of these measures are given below.

Positive and negative predictive values

Diagnostic process concentrates on the probability whether a person is diseased or not rather than sensitivity and specificity values of the test. These probabilities known as post-test may be more guiding at the implementation stage of the test.

Positive Predictive Value (PPV) is the probability of disease in a subject with positive test result, and defined as the predictive value of positive test result. Here, P(D+|T+) denotes the probability that the subject concerned is actually diseased when the subject's test result is known to be positive.

$$PPV = P(D+|T+) = TP/(TP + FP)$$

Negative Predictive Value (NPV) is the probability that a subject with negative test result is not diseased and shown as:

$$NPV = P(H-|T-) = TN/(FN + TN)$$

Since PPV and NPV are both affected by the prevalence of the disease or by its pre-test probability, tests with same levels of sensitivity and specificity may yield different PPV and NPV values in groups with different prior probabilities [12]. Hence, while assessing test results, prior probability (prevalence) of the disease in test groups must be considered.

Prevalence-dependent PPV and NPV values can be calculated successively as follows:

$$PPV = P(D+|T+) = \frac{\text{Sensitivity} \times \text{Prevalence}}{\text{Sensitivity} \times \text{Prevalence} + (1 - \text{Specificity}) \times (1 - \text{Prevalence})}$$

$$NPV = P(D-|T-) = \frac{\text{Specificity} \times (1 - \text{Prevalence})}{\text{Specificity} \times (1 - \text{Prevalence}) + (1 - \text{Sensitivity}) \times \text{Prevalence}}$$

Diagnostic odds ratio (DOR)

Diagnostic odds ratio is one of the measures giving the overall performance of a diagnostic test. In diagnostic tests, odds ratio is defined as the ratio of the odds of the test being positive among the diseased relative to the odds of the test being positive among the healthy. It is a measure not affected by prevalence.

$$DOR = \frac{\text{Sensitivity} \times \text{Specificity}}{\text{False Negative rate} \times \text{False Positive Rate}}$$

Since the odds ratio is not a probability value it ranges from zero to infinity. It shows the likelihood of a test to yield positive result for the diseased relative to the healthy. As the odds ratio gets greater and greater than 1 it means that the discriminative power of the test is also higher.

- OR>1 indicates that the positive test result among patients is more likely than healthy.
- OR=1 indicates that the test does not contribute to discrimination where the true positive ratio equals the false positive ratio (TPR=FPR),
- OR<1 indicates that the test has a worse discrimination than chance. It is not expected to be observed in practice.

Youden Index (YI)

This measure gives an overall value for the performance of a diagnostic test. While mostly used to determine the overall performance of a test, YI can also be used in comparing more than one diagnostic test.

It is an indicator how greater is the likelihood of positive test result among the diseased than among the healthy. The Youden index ranges values from 0 to 1 ($0 \leq YI \leq 1$). Getting closer to 0 means test has low performance in discrimination and the opposite as it approaches to 1. The result may turn out as negative in case FP is greater than TP which cannot be expected in practice since the test developed must not be worse than mere chance.

$$YI = \text{Sensitivity} + \text{Specificity} - 1 = TPR - FPR$$

The Youden Index is also one of the measures used to determine the optimum cut-off point in tests that yield continuous numerical results. Particularly in cases where the researcher assigns equal importance to positive and negative classification performance in a test, the highest point in the Youden Index is used as a criterion to determine the optimum cut-off point.

Likelihood ratio (LR)

The likelihood ratio is a measure showing the performance of tests by using sensitivity and specificity values together. There are two different measures as positive likelihood ratio and negative likelihood ratio [13].

Positive likelihood ratio (LR+)

It is the ratio of the probability of obtaining positive test results in patients to the probability of obtaining positive results in healthy subject.

$$LR(+) = \frac{P(T+|D+)}{P(T+|D-)} = \frac{\text{Sensitivity}}{1 - \text{Specificity}}$$

$$LR(+) = \frac{TPR}{FPR}$$

The positive likelihood ratio is the best indicator in ruling in the presence of the disease. Like the odds ratio, the value of this ratio ranges from 1 to infinity. It is possible for this measure taking values between 0 and 1 values also, but this value is not expected to be worse than chance. The higher the LR(+) value than 1, the more successful the positive results of the index test.

It shows how many times it is possible for a test to give positive result in persons with disease relative to the same result in healthy persons. For example if LR(+) value is greater than 10 it means that the discrimination capacity is high and thus there is great change from prior probability to posterior disease probability (PPV). When LR(+) is equal to 1 this means TPR=FPR which indicates that the test is not informative beyond mere chance.

Negative likelihood ratio (LR−)

It is the ratio of the probability of obtaining negative test results in patients to the probability of obtaining negative results in healthy subjects.

It is defined as follows:

$$LR(-) = \frac{(T - |D+)}{(T - |D-)} = \frac{(1 - \text{Sensitivity})}{\text{Specificity}}$$

$$LR(-) = \frac{FNR}{TNR}$$

LR(–) is a good indicator for ruling out the disease. As the ratio gets smaller than 1 the test is considered as successful in negative results and the ratio closer to 1 means that the test is not successful. In other words, FNR=TNR and it means that that the test is useless. The value of this ratio ranges from 0 to 1. LR(–) <0.1 is the indicator of a good test.

Further, the likelihood ratio has advantages like being a measure independent of disease prevalence and easily obtaining posterior probabilities from prior probabilities with the help of LR. Necessary equations can be found in references [13, 14].

Fagan's nomogram is alternative method for calculating post-test probability. Post-test probability can be obtained easily from Fagan's nomogram which is one simple method. It is a graphical method used for calculating post-test probability, knowing pre-test probability and likelihood ratio. It can be accessed from many articles [13–15].

As likelihood ratio changes there is also change in posterior probabilities together with change in prior probabilities. This change can be shown as in Figure 1 below.

As LR(+) value increases the curve gets closer to the upper left corner. As can be gathered from the figure, posterior probabilities increase as prior probabilities increase. As test performance (LR (+)) increases there is higher increase in posterior probabilities. However, when LR(+)=1 (diagonal line) posterior probability takes the same value as priori probability and it is observed that the test does not cause any change.

As the curve gets closer to the upper left corner it is observed that tests classification performance gets better. For example, when LR(+)=1 in a situation where prior probability is 10%, posterior probability becomes 10%, and posterior probability increases to 70% when LR(+) is 20. Although the prior probability remains the same, it can be seen that the contribution of the test to the posterior probability (PPV) is increasing with the increase in LR(+).

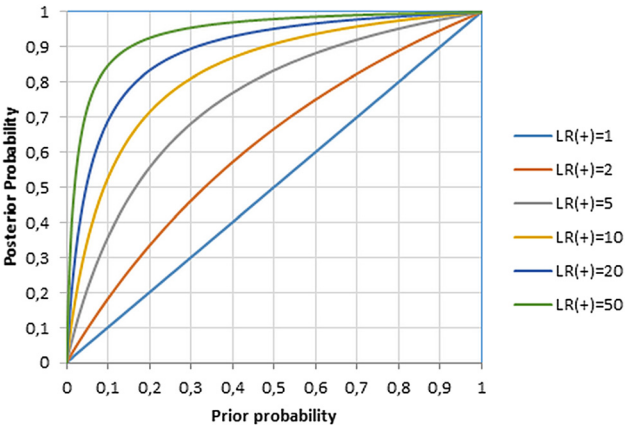

Figure 1: Relationship between prior probability and posterior probability as likelihood ratio increases.

ROC analysis

When the diagnostic test result is numerical, ordinal or binay, the performance of the diagnostic test is examined by Receiver Operating Characteristic (ROC) Analysis. However, this method is often used for tests with numerical results. This analysis is a very comprehensive topic, so it has been mentioned very briefly in this review.

Receiver Operating Characteristic (ROC) Curve is drawn to evaluate the test performance. Area under the ROC curve (AUC) provides an overall measure of the index test performance. There are many types of ROC analysis. When the gold standard test has two categories (e.g. patient-healthy), two-way ROC analysis, when it has three categories (e.g. patient-risky-healthy) three-way ROC analysis, and when it has more than three categories (Stage I, Stage II, Stage III, Stage IV), multi-class ROC analysis is used. ROC Analysis can be performed even for the gold standard test result to be numerical. Two-way ROC analysis is the most widely known and applied method. This method can be applied in many different statistical softwares (IBM SPSS Statistics, MedCalc, Stata, SAS, R etc.). However, it is not possible to perform other ROC analyzes in every program. For these, some packages written in R software or codes written in other programs are used.

Confidence intervals in test performance measures

It is practically impossible to conduct a research on the whole population of interest due to reasons such as cost, time needed, etc. Studies are therefore conducted by selecting a sample from the population concerned which represents its population with similar characteristics. We try to estimate unknown characteristics of the population (population parameters) by working on the sample. In other words, a sample is randomly selected from the target

population and sample statistics are used to estimate unknown population parameters. Statistics calculated using a single sample is the point estimate of population parameter. In addition to point estimation in statistics, it is more appropriate to give an interval estimation that indicates the values that the unknown population parameter can take at a certain level of confidence (though not a rule, this level is often selected as 95%). A Confidence interval is a range of values that likely would contain an unknown population parameter [2, 16].

The researcher has the chance of obtaining different measures from different samples that can be drawn from the same population. This variation of measures from sample to sample is defined as standard error (SE). Confidence interval is calculated with the help of standard error. Confidence interval has its upper and lower limits. The lower limit is obtained by multiplying the value obtained from theoretical statistical distributions for the desired confidence level (which is $Z_{1-0.05/2}$=1.96 for a confidence level of 95%) with standard error and subtracting this value from point estimation, and the upper limit by adding this value to point estimation. As the value for standard error increases confidence interval gets larger and vice versa. Since standard error will get smaller as sample size gets larger, the resulting confidence interval will be narrower.

When performance measures are calculated, it is better to present these values together with confidence intervals. This interval shows the probability in percentage terms that unknown actual population value falls in the interval determined on the basis of the sample. For example, selection of the level 95% means the unknown population parameter will have a value within the calculated confidence interval with the probability of 95%. The margin of error in this case will be 1-confidence level=1–0.95=0.05 (alpha) or 5%. This means that the unknown population parameter may remain out of the interval with the probability of 0.05 (α).

Confidence intervals can be calculated for all test performance measures. Confidence interval estimation for the proportion is used while computing confidence intervals for all performance measures (sensitivity, specificity, PPV, NPV, overall accuracy, Youden Index) ranging from 0 to 1. Lower and upper limits for an asymptotic confidence interval are obtained with the help of the following equation:

$$\text{Confidence Interval (CI)} = \hat{p} \pm Z_{1-\alpha/2} \times \text{SE}(\hat{p})$$

$$\text{SE}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$\text{Lower limit (LL)} = \hat{p} - Z_{1-\alpha/2} \times \text{SE}(\hat{p})$$

$$\text{Upper limit (UL)} = \hat{p} + Z_{1-\alpha/2} \times \text{SE}(\hat{p})$$

Here "p" can be thought of any of the test performance measures expressed as percentage terms. One can find different methods of confidence interval estimates in the literature [1,2]. Differences may be associated with the methods used.

Confidence interval estimates are also obtained for odds ratio and likelihood ratio measures. In both of these measures, obtaining the value 1 means that the test was not informative. Thus, measures with confidence intervals including 1 denote measures that have no statistical significance. Sample distributions of odds ratio and likelihood ratio are not symmetrical [17]. However, logarithms of both measures display approximately normal distribution. Asymptotic confidence interval may be used as a method after taking logarithms. After taking the natural logarithm (ln) of odds ratio confidence interval is obtained as in the equation:

$$\ln\left(\widehat{OR}\right) \pm Z_{1-\alpha/2} \times \text{SE}\left(\ln\left(\widehat{OR}\right)\right)$$

But here, it is more appropriate to present the results after transforming the confidence interval obtained into its original unit by taking the exponential.

$$\text{CI}\left(\widehat{OR}\right) : e^{\ln\left(\widehat{OR}\right) \pm Z_{1-\alpha/2} x \text{SE}\left(\ln\left(\widehat{OR}\right)\right)}$$

The standard error here is:

$$\text{SE}\left(\ln\left(\widehat{OR}\right)\right) = \sqrt{\left(\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}\right)}$$

The confidence interval estimate for likelihood ratio is made as similar to OR by taking the natural logarithm of LR(+):

$$\text{CI}\left(\widehat{LR(+)}\right) : \exp\left(\ln\left(\widehat{\text{LR}(+)}\right) \pm Z_{1-\alpha/2} \times \sqrt{\frac{1-\widehat{\text{Sens}}}{TP}+\frac{\widehat{\text{Spec}}}{FP}}\right)$$

$$\text{CI}\left(\widehat{LR(+)}\right) : \widehat{LR(+)} \times e^{\pm Z_{1-\alpha/2} \times \sqrt{\frac{1-\widehat{\text{Sens}}}{TP}+\frac{\widehat{\text{Spec}}}{FP}}}$$

where "Sens" is the sensitivity and "Spec" is the specificity.

The confidence interval estimate for LR(–) can be obtained by the following equation:

$$\exp\left(\ln\left(\widehat{\text{LR}(-)}\right) \pm Z_{1-\alpha/2} \times \sqrt{\frac{\widehat{\text{Sens}}}{FN}+\frac{1-\widehat{\text{Spec}}}{TN}}\right)$$

$$\mathrm{CI}\left(LR(-)\right): \widehat{LR(-)} \times e^{\pm Z_{1-\alpha/2}\times\sqrt{\frac{\widehat{\mathrm{Sens}}}{TN}\frac{1-\widehat{\mathrm{Spec}}}{TN}}}$$

Instead of manual calculation one can easily reach confidence intervals with the help of some statistical softwares or calculators available in the web [18, 19].

Sample size calculation for sensitivity and specificity

It is very important to determine properly sample size according to the study design and the objective of the study. The main aim of a sample size calculation is to determine the number of participants needed to detect a clinically relevant effect. When the sample size is chosen too small, one may not be detected a clinically important effect, whereas sample size is chosen too large one may waste time, resources, etc. Therefore, it is important to calculate the optimum sample size.

When the true disease status of subjects' is known, the sample size can be calculated as in the following equation to estimate the sensitivity or specificity values of a new diagnostic test.

$$n = \frac{Z^2_{\alpha/2}\widehat{P}\left(1-\widehat{P}\right)}{d^2}$$

"P" is a proportion value as either sensitivity or specificity. $\widehat{P}$ is pre-determined value, d is the precision of estimate (i.e. the maximum marginal error). Z value for α=0.05 is used as 1.96.

Generally, true disease status is unknown at the time of sampling. Also, clinicians often want to estimate both sensitivity and specificity at the same time. Because of that, while calculating the total sample size, the prevalence of disease needs to be included in calculation [20].

In this case, when the prevalence (Prev) is taken into account, the sample size calculation is obtained as in the following equations [21].

$$n_{\mathrm{Sens}} = \frac{Z^2_{\alpha/2}\widehat{\mathrm{Sens}}\left(1-\widehat{\mathrm{Sens}}\right)}{d^2\mathrm{Prev}}$$

$$n_{\mathrm{Spec}} = \frac{Z^2_{\alpha/2}\widehat{\mathrm{Spec}}\left(1-\widehat{\mathrm{Spec}}\right)}{d^2\left(1-\mathrm{Prev}\right)}$$

Final sample size (N) is the larger of n_{sens}, n_{spec} [20]. Therefore, the sample size will be adequate for estimation of both sensitivity and specificity with desired precision by this way.

Table 2. Sample size calculation formulas in studies of test accuracy.

Purpose	Formula
Sample size to estimate both sensitivity and specificity and their confidence intervals	$n_{\mathrm{Sens}} = \frac{Z^2_{\frac{\alpha}{2}}\widehat{\mathrm{Sens}}(1-\widehat{\mathrm{Sens}})}{d^2\mathrm{Prev}}$ $n_{\mathrm{Spec}} = \frac{Z^2_{\frac{\alpha}{2}}\widehat{\mathrm{Spec}}(1-\widehat{\mathrm{Spec}})}{d^2(1-\mathrm{Prev})}$ Sens or Spec: pre-determined values of sensitivity or specificity Prev: predetermined valu of prevalence of disease $Z_{a/2}$: for a=0.05, $Z_{0.05/2}$ = 1.96 d: the precision of estimate or the maximum marginal error Final sample size(N) is the larger of n_{sens}, n_{spec} for estimation of both sensitivity and specificity
Sample size for testing sensitivity or specificity of single test Hypothesis: H_0: Sens=P_0 H_1: Sens≠P_0 (or Sens=P_1) or H_0: Spec=P_0 H_1:Spec≠P_0 (or Spec=P_1)	$n = \frac{\left[Z_{\frac{\alpha}{2}}\sqrt{P_0(1-P_0)}+Z_\beta\sqrt{P_1(1-P_1)}\right]^2}{(P_1-P_0)^2}$ P_0: predetermined value of sensitivity or specificity of new diagnostic test P_1: the value of sensitivity or specificity under alternative hypothesis. $Z_{\alpha/2}$: for α=0.05, $Z_{0.05/2}$= 1.96 Z_β: for β=0.20, Z=0.84
Sample size for comparing the sensitivity or specificity of two diagnostic tests Hypothesis: H_0: P_1=P_2 H_1: P_1≠P_2	$n = \frac{\left[Z_{\frac{\alpha}{2}}\sqrt{2\bar{n}\bar{P}(1-\bar{P})}+Z_\beta\sqrt{P_1(1-P_1)+P_2(1-P_2)}\right]^2}{(P_1-P_2)^2}$ $\bar{P}$: average of P_1 and P_2 P_1 or P_2: expected sensitivity or specificity of two diagnostic tests $Z_{\alpha/2}$: for α=0.05, $Z_{0.05/2}$= 1.96 Z_β: for β=0.20, Z=0.84

Sample sizes have been calculated to estimate sensitivity and specificity for different margin values and prevalence values in some articles [20, 21]. Tables have been prepared. Researchers can use these tables for their own studies. Some calculators are also available for these simple calculations on the web.

The aim of a study may not be just to estimate a performance measure. Instead, it may be desirable to test the performance of a new diagnostic test by hypothesis testing against a specific value or to compare the performance (sensitivity or specificity) of the two diagnostic tests [22, 23]. Equations can be used in calculating the samples size for these purposes are summarized in Table 2. Sample size can be calculated for comparison of single proportion, two proportions or two areas under the ROC curves by using PASS or MedCalc softwares.

Application

Coronaviruses (CoV) constitute a family whose manifestations range from mild infections to more serious and even fatal ones like Middle East Respiratory Syndrome (MERS) and Severe Acute Respiratory Syndrome (SARS). The new coronavirus emerging in 2019 and known as SARS CoV 2 causes the disease COVID-19. Accurate and quick diagnosis is as important as treatment in this infection. Early and accurate diagnosis makes it possible to isolate individuals with positive test result and start early treatment.

The method reverse transcription-polymerase chain reaction (RT-PCR) is used in diagnosing the Covid 19 disease. The Polymerase Chain Reaction (PCR) is a molecular test for screening the RNA of the new coronavirus in specimen collected from nasopharyngeal airway or sputum. It is based on the confirmation of RNA series of the virus and recognized as the best in Covid-19 diagnosis. The time required for the result may vary, but it takes at least 4 h. Besides, there are also some rapid test kits developed. Rapid diagnosis tests screen virus antigens or antibodies developed by the human immune system. Rapid diagnosis tests look for virus antigen in nasal swab. Antibodies developed against the virus are checked in blood. In both PCR and antigen tests the result for Covid-19 is binary as either negative or positive [23].

The accuracy of results in PCR and Antigen tests may vary depending on various conditions including their application in too early or too late phases of the infection, poor collection or storage. Reserving for these, the PCR test is still considered as the best and reference one.

Many studies have been conducted to investigate the performance of rapid tests according to the RT-PCR test for Covid 19 [24, 25]. A hypothetical data has been produced based on the studies conducted on this subject in the literature. The data generated to examine the performance of the immunochromatography (Specific IgM/IgG) vs. RT-PCR are given Table 3.

All performance measures and their confidence intervals for the data given in the table were calculated. To obtain these measures it is sufficient to enter TP, FP, FN and TN numbers to four cells in calculators found in the web. All performance measures and their confidence intervals that can be used for diagnostic tests with binary results are given in Table 4.

Table 3: Diagnostic outcome of Specific IgM/IgG vs. RT- PCR reference standard test.

Specific IgG/IgM (index test)	RT-PCR test (reference test)		Total
	Disease present (D+)	Disease absent (D−)	
Test positive (T+)	203 (TP)	10 (FP)	213
Test negative (T−)	88 (FN)	149 (TN)	237
Total	291	159	450

Table 4: Diagnostic performance value and 95% confidence interval of Specific IgM/IgG vs. RT- PCR reference standard test.

Statistics (indicator)	Equation	Measures of diagnostic test performance	95% CI (LL–UL)
Sensitivity (TPR)	TP/(TP + FN)	69.76%	64.13–74.98%
Specificity (TNR)	TN/(FP + TN)	93.71%	88.74–96.94%
PPV	TP/(TP + FP)	95.31%	91.73- 97.38%
NPV	TN/(FN + TN)	62.87%	58.60–66.95%
LR(+)	Sensitivity/(1 − specificity)	11.09	6.06–20.31
LR(−)	(1 − Sensitivity)/Specificity	0.32	0.27–0.39
DOR	(Sensitivity × Specificity)/(FNR × FPR)	34.37	17.29–68.35
Accuracy	(TP + TN)/(TP + TN + FP + FN)	78.22%	74.12–81.95%
Disease prevalance	(TP + FN)/(TP + TN + FP + FN)	64.67%	60.05–69.09%

The results in this table are obtained from MedCalc's free online "Diagnostic test statistical calculator" [18]. Apart from these measurements, another measurement for binary diagnostic test is the Youden index. YI for this table was calculated as approximately YI=0.6976 + 0.9371–1=0.64

As can be seen in the table, the approximate specificity value is 94% and sensitivity value is 70% in this 2 × 2 contingency table. While performance measure in specificity is quite high, the sensitivity is not as high as specificity. It can be said that positive results of the test are more reliable than its negative results which can also be understood from higher PPV value than NPV. Since the prevalence of the disease is unknown, PPV and NPV value was obtained without using the Bayes theorem in the table.

The test's likelihood of giving positive results for the diseased is 11.09 times greater than non-diseased. Since confidence intervals for values of LR(+), LR(–) and DOR do not include 1 it can be said that it is statistically significant. Some calculation tools may yield difference results in estimating confidence intervals which may derive from different standard error and confidence interval estimates.

Conclusions

It is important to design and evaluate the performance of diagnostic test or screening test for health care. False positive and false negative results are not obtained only in a perfect test. In other tests, patients and healthy individuals cannot be completely separated from each other, and there may be misclassifications. Therefore, the accuracy of these tests needs to be assessed. For diagnostic tests with binary results, test performance measures are evaluated using sensitivity, specificity, predictive values, overall accuracy, Youden index, diagnostic odds ratio, and likelihood ratio. Each measure brings different interpretations to the performance of the test. Some measures are affected by prevalence, while others are not. For predictive values affected by the prevalence, it is necessary to make predictions that also take into account the prevalence value. In studies, it is useful to present performance measures with their confidence intervals.

Accurately predicting the performance of a diagnostic test depends on many factors. These factors can be study design (cross-sectional, retrospective or prospective), whether participants formed a consecutive or random, rationale for choosing the reference standard (if alternatives exist), whether clinical information and reference standard results were available to the performers/readers of the index test, sample size etc. There are guidelines that ensure that all information regarding the conditions under which the study was conducted is in the report and that the method of the study is written more accurately, in terms of such factors. One of the most important guidelines developed for diagnostic accuracy studies is STARD. STARD, which was first created in 2003 and stands for "Standards for Reporting of Diagnostic Accuracy Studies", is a guideline developed in order to increase the quality and set a standard in the reporting of diagnostic accuracy studies [26]. This guideline has the latest updated version of the criteria consisting of 30 items in 2015. Since this checklist ensure that a report of a diagnostic accuracy study contains the necessary information, it is recommended for use by many publishers.

Research funding: None declared.

Author contribution: The author has accepted responsibility for the entire content of this submitted manuscript and approved submission.

Competing interests: The authors declare no conflicts of interest regarding this article.

References

1. Zhou XH, Obuchowski NA, McClish DK. Statistical methods in diagnostic medicine. New York: John Wiley & Sons; 2002: 15–40 pp.
2. Karaağaoğlu E, Karakaya J, Kılıçkap M. Tanı testlerinin değerlendirilmesinde istatistiksel yöntemler. Ankara: Detay Yayıncılık; 2016:1–57 pp.
3. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982; 143:29–36.
4. Mandrekar JN. Receiver operating characteristic curve in diagnostic test assessment. J Thorac Oncol 2010;5:1315–6.
5. Hoffmann T, Bennett S, Del Mar C. Evidence-based practice across the health professions. Australia: Elsevier; 2017:20–30 p.
6. Knottnerus JA, Weel C, Muris J. Evidence base of clinical diagnosis evaluation of diagnostic procedures. BMJ 2002;324:477–80.
7. Parikh R, Mathai A, Parikh S, Sekhar GC, Thomas R. Understanding and using sensitivity, specificity and predictive values. Indian J Ophthalmol 2008;56:45–50.
8. Detrano R, Janosi A, Lyons KP, Marcondes G, Abbassi N, Froelicher VF. Factors affecting sensitivity and specificity of a diagnostic test: the exercise thallium scintigram. Am J Med 1988; 84:699–710.
9. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. N Engl J Med 1978;299: 926–30.
10. Whiting P, Rutjes AW, Reitsma JB, Glas AS, Bossuyt PM, Kleijnen J. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. Ann Intern Med 2004;140:189–202.
11. Karakaya J, Aksoy DY, Harmanci A, Karaaagaoglu E, Gurlek A. Predictive ability of fasting plasma glucose for a diabetic 2-h postload glucose value in oral glucose tolerance test: spectrum effect. J Diabetes Complicat 2007;21:300–5.

12. Altman DG, Bland JM. Diagnostic tests 2: predictive values. BMJ 1994;309:102.
13. Florkowski CM. Sensitivity, specificity, receiver-operating characteristic (ROC) curves and likelihood ratios: communicating the performance of diagnostic tests. Clin Biochem Rev 2008;29(1 Suppl):83–7.
14. Deeks JJ, Altman DG. Diagnostic tests 4: likelihood ratios. BMJ 2004;329:168–9.
15. Fagan TJ. Nomogram for Bayes's theorem. N Engl J Med 1975;293: 257.
16. Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. Br Med J 1986; 292:746–50.
17. Bland M, Altman D. Statistics notes. The odds ratio. Br Med J 2000;320:1468.
18. MedCalc Statistical Software. Free statistical calculator. Available from: https://www.medcalc.org/calc/diagnostic_test.php [Accessed 04 Jul 2020].
19. Stats V. Clinical calculator 1. Available from: https://vassarstats.net/clin1.html [Accessed 04 Jul 2020].
20. Buderer NM. Statistical methodology: incorporating the prevalance of disease into the sample size calculation for sensitivity and specificity. Acad Emerg Med 1996;3:895–900.
21. Karimollah H-T. Sample size estimation in diagnostic test studies of biomedical informatics. J Biomed Inf 2014;48:193–204.
22. Obuchowski NA. Sample size calculations in studies of test accuracy. Stat Methods Med Res 1998;7:371–92. .
23. Carter LJ, Garner LV, Smoot JW, Li Y, Zhou Q, Saveson CJ, et al. Assay techniques and test development for COVID-19 diagnosis. ACS Cent Sci 2020;6:591–605.
24. Hoffman T, Nissen K, Krambrich J, Rönnberg B, Akaberi D, Esmaeilzadeh M, et al. Evaluation of a COVID-19 IgM and IgG rapid test; an efficient tool for assessment of past exposure to SARS-CoV-2. Infect Ecol Epidemiol 2020;10:1754538.
25. Bastos ML, Tavaziva G, Abidi SK, Campbell JR, Haraoui LP, Johnston JC, et al. Diagnostic accuracy of serological tests for covid-19: systematic review and meta-analysis. BMJ 2020;370:m2516.
26. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. BMJ 2015;351:h5527.