Cardiovasc Prev Pharmacother. 2021 Jan;3(1):15-20
https://doi.org/10.36011/cpp.2021.3.e2
eISSN 2671-700X

Cardiovascular Prevention
and Pharmacotherapy

Special Article

Methods for Evaluating the Accuracy of Diagnostic Tests

Chi-Yeon Lim, PhD

Department of Biostatistics, Dongguk University College of Medicine, Goyang, Korea

OPEN ACCESS

Received: Dec 16, 2020
Accepted: Jan 17, 2021

Correspondence to
Chi-Yeon Lim, PhD
Department of Biostatistics, Dongguk University College of Medicine, 32 Dongguk-ro, Ilsandong-gu, Goyang 10326, Korea.
E-mail: rachun@hanmail.net

ORCID iDs
Chi-Yeon Lim
https://orcid.org/0000-0003-0178-6976

Conflict of Interest
The author has no financial conflicts of interest.

ABSTRACT

The accuracy of a diagnostic test should be evaluated before it is used in clinical situations. The sensitivity, specificity, and the trade-off between the 2 need to be considered. Sensitivity and specificity in diagnostic tests depend on the selection of cut-off values, and appropriate cut-off values can be arrived at by analyzing the receiver operating characteristic curve. In actual clinical setting, it is often difficult to obtain an appropriate gold standard for diagnosis, and in this case, consent is required as well. In this article, we summarize the basic concepts and methods for evaluating the performance of diagnostic tests.

Keywords: Biostatistics; Causality; Epidemiologic studies; Sensitivity and specificity; Statistics

INTRODUCTION

With the rapid spread of the severe acute respiratory syndrome coronavirus 2, the Food and Drug Administration has reported that the pandemic is disrupting healthcare system and social stability worldwide and that rapid detection of cases and contacts, appropriate clinical management and infection control, and implementation of community mitigation efforts are necessary to effectively respond to coronavirus disease 2019 outbreak. In the context of a public health emergency, all clinical trials or diagnostic studies should be validated before use. It is very important to verify diagnostic accuracy because false results can negatively affect not only individual patient's health but also public health in general.[1)]

Diagnostic tests play an important role in identifying the patient's disease or condition, and the accuracy of diagnostic tests must be determined using comparisons and statistical tests. There is a screening test involved in the process of identifying the diseases that are present. It is important not to miss the individuals suffering from the disease when conducting the screening test. Therefore, the sensitivity must be increased. It would be ideal to have both high sensitivity and high specificity, but in the same test, higher sensitivity would result in lower specificity. Thus, there is a trade-off between sensitivity and specificity that affects the accuracy of diagnosis in diagnostic studies.

There are several methods for evaluating the performance of a diagnostic test. The sensitivity and specificity of the test are well known parameters with known disease status based on the gold standard. Toft et al. explored methods using simulations without latent class analysis to assess the sensitivity and specificity of a diagnostic test.[2] It can be problematic to evaluate the accuracy of diagnostic test because there is no "gold standard" for diagnostic testing in many fields of medical research.[3] For a diagnostic test, sensitivity and specificity vary with the choice of cut-off value, and the analysis of receiver operating characteristic (ROC) curve can help one arrive at an appropriate cut-off value.[4]

It can estimate some statistics for accuracy, precision, and recall by defining "true disease" or using "gold standard." If there are no pre-informative results using "true disease" or "gold standard," it is hard to confirm accuracy, although estimating an agreement is still possible. The purpose of this review is to introduce the basic concepts and methods of performance to evaluate the accuracy of a diagnostic test.

TYPES OF TEST RESULTS

There are various measures of the accuracy of diagnostic tests, and the choice of method of comparison depends on the nature of the test results. The results can be classified as either qualitative, such as binary-scale or ordinal-scale data, or quantitative, such as continuous-scale data. This review considers the simplest type of dichotomous test with only 2 possible results (positive and negative). There are 4 subgroups of diagnostic results of patients: D+ means patients who have the disease, D− means without the disease, and T+ and T− mean test positive and negative, respectively.

Gold standard

A gold standard is important for decision-making by the medical staff when it comes to applying evidence-based medicine.[5] It is also known as the "reference standard," and it is considered to be the best available method for evaluating the presence or absence of the target condition (or disease). In other words, it is a source of information that is completely different from the test or tests under evaluation and shows the true condition status (D+, D−) of the patient. The selection of the gold standard is often the most difficult step in the planning phase of the study. If it exists, the study for diagnostic accuracy tests is easier to carry out. However, there is no reasonable gold standard available at present, and some argue that this issue should be resolved before beginning the study.

Bias

It is well known that diagnostic accuracy tests are subject to biases of various types, such as verification, errors in the reference spectrum, and test interpretation bias. Estimates of the diagnostic performance, including sensitivity and specificity, can also be biased. Biased estimates do not provide the true sensitivity and specificity values, thereby causing inaccuracy. There has been some debate about bias in diagnostic tests.[5]

Definition of measures for diagnostic tests

There are 2 basic measures of diagnostic accuracy: sensitivity and specificity. In **Table 1**, the rows summarize the test results, whereas the columns illustrate the data according to the status of true condition using the gold standard or standard reference. Other important definitions are included in the footnote of the table.

Table 1. A 2×2 count table for comparing a new test result to a reference standard (using gold standard)

Test result	Condition (disease)		Total
	Positive (D+)	Negative (D−)	
Positive (T+)	n_{11}	n_{12}	$n_{1.}$
Negative (T−)	n_{21}	n_{22}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	N

Data are represented as below: n_{11}, number of true positive events; n_{12}, number of false positive events; n_{21}, number of false negative events; n_{22}, number of true negative events.

Sensitivity and specificity are the parameters for estimating how often the test result is positive when the target condition is present (+) and how often the test is negative when the target condition is absent (−), respectively. When both estimates are close to 1, they perform well in terms of diagnostic ability. Most results are reported in a 2×2 table, as exemplified by **Table 1**. Estimated sensitivity and specificity are the proportions of subjects with and without the condition of the reference standard (or gold standard), respectively. Predictive values (positive predictive value [PPV], negative predictive value [NPV]) define the probability of being ill for a subject with a positive or negative result. Likelihood ratios (LR+, LR−) are the ratios of sensitivity to specificity.[6] From **Table 1**, some estimates, including sensitivity and specificity, are calculated as follows.

$$\text{Sensitivity (Se)}: \frac{n_{11}}{n_{.1}} \text{ (true positive rate)}$$

$$\text{Specificity (Sp)}: \frac{n_{22}}{n_{.2}} \text{ (true negative rate)}$$

$$\text{Accuracy: } \frac{n_{11}+n_{22}}{N} = \frac{\text{Se}+\text{Sp}}{N}$$

$$\text{PPV: } \frac{n_{11}}{n_{1.}}$$

$$\text{NPV: } \frac{n_{22}}{n_{2.}}$$

$$\text{LR}+: \frac{Se}{1-Sp}$$

$$\text{LR}-: \frac{1-Se}{Sp}$$

Agreement

When a non-reference standard is used for evaluating a new diagnostic test, sensitivity and specificity are no longer adequate estimates. In this case, agreement is used for the comparison of information on the correctness. In a 2×2 table for comparing a new test result to a non-reference standard (**Table 2**),

$$\text{Overall percent agreement: } \frac{n_{11}+n_{22}}{N}$$

$$\text{Positive percent agreement (PPA)}: \frac{n_{11}}{n_{.1}}$$

$$\text{Negative percent agreement (NPA)}: \frac{n_{22}}{n_{.2}}$$

Table 2. A 2×2 count table for comparing a new test result to a non-reference standard

Test result	Non-reference standard		Total
	Positive (D+)	Negative (D−)	
Positive(T+)	n_{11}	n_{12}	$n_{1\cdot}$
Negative(T−)	n_{21}	n_{22}	$n_{2\cdot}$
Total	$n_{\cdot 1}$	$n_{\cdot 2}$	N

Data are represented as below: n_{11}, number of true positive events; n_{12}, number of false positive events; n_{21}, number of false negative events; n_{22}, number of true negative events.

The counts in the table (n_{11}, n_{12}, n_{21}, and n_{22}) no longer represent true positive, false positive, false negative, and true negative because the non-reference standard may be wrong. They only provide information on how often the test results agree with a non-reference standard. In this case, PPA and NPA are more useful in describing their agreement.[7]

ROC curve

In the ROC curve with specificity on the x-axis and sensitivity on the y-axis, the area under the curve (AUC) is useful for evaluating how high the discriminative power of the test is. The closer the AUC value is to 1, the higher the diagnostic accuracy. If it is less than 0.5, then the diagnostic accuracy is low.

Trade-off between sensitivity and specificity

When choosing for a positive diagnosis, there is a trade-off between sensitivity and specificity depending on the cut-off value.[8] Unfortunately, it is difficult to have both 100% sensitivity and 100% specificity when deciding the cut-off value. There is no mathematical relationship between sensitivity and specificity; however, as one increases, the other tends to decrease. Therefore, achieving the optimal balance between the 2 is very important. To achieve higher sensitivity, a trade-off must occur between the 2. In **Figure 1**, the true and false positive values are to the right of the dotted line since this distinguishes "without disease" from "with disease." High cut-off values make diagnostic tests highly specific but reduce sensitivity.

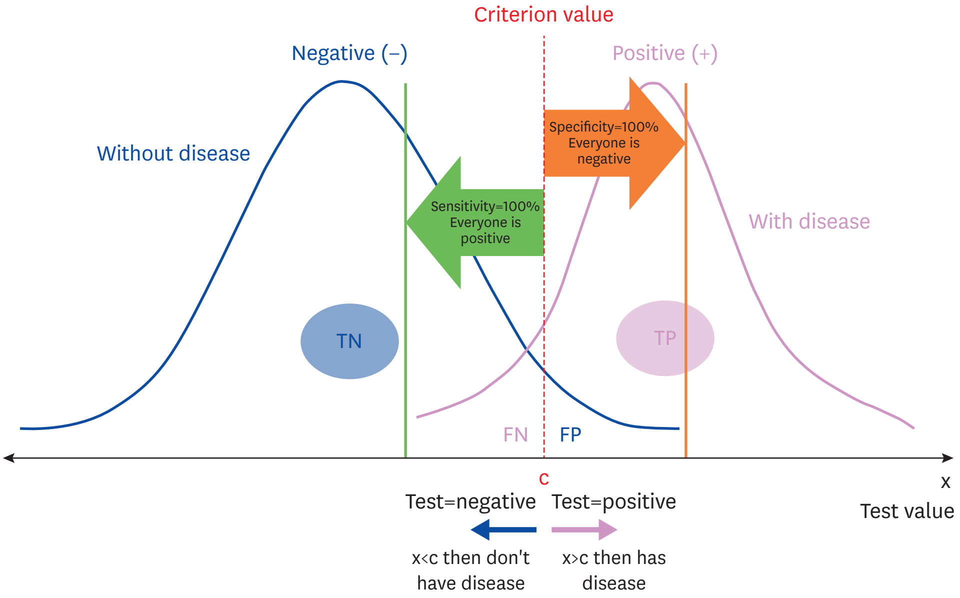

Figure 1. The trade-off between sensitivity and specificity.
FN = false negative; FP = false positive; TN = true negative; TP = true positive.

In contrast, low cut-off values for disease mean that more test cases are considered "with disease"; therefore, the test evaluates more diseases.

Sample size estimation for diagnostic test

In diagnostic research, sample size calculation plays an important role in ensuring validity and reliability. The right sample size is calculated based on the study objective. There are several objectives of a diagnostic study for accuracy, including the following: 1) estimating the accuracy of a diagnostic test, 2) determining whether 2 diagnostic tests have different accuracies, 3) evaluating whether 2 different diagnostic tests have equivalent accuracies,[9] and 4) identifying an appropriate cut-off value for the test. For the first objective, the sample size must be calculated to ensure that the accuracy is estimated to a pre-specified precision. For the second objective, the sample size must be large enough to test whether the accuracies of the 2 tests are different, and the study must prove that they have a high enough power to detect differences. For the third objective, the sample size needs to be large enough to ensure that if the accuracies are truly equivalent, the study will have a high enough power for determining the equivalence. For the fourth objective, it should be ensured that the chosen cut-off value reaches the minimum requirements for sensitivity and specificity. The sample size for diagnostic accuracy studies is calculated based on the measures of accuracy, such as sensitivity, specificity, ROC curve, and sensitivity at a fixed false positive rate.

LIMITATIONS

There may not exist a consensus reference standard, or the reference standard might be error prone because of the non-negligible percentage of the population in need. In these cases, it would be better to consult before planning the study on the choice of reference standard or statistical methods. To reduce the bias of estimates of sensitivity and specificity for diagnostic research, simply increasing the total sample size for the study will not be helpful. However, choosing the "right" subject during the planning phase, designing to eliminate bias in the study, and performing the "right" data analysis procedures are more helpful in eliminating or reducing bias.

CONCLUSION

The first step in a diagnostic study for accuracy is to set up the primary objective. When the objective is clearly stated, "right" subjects can be selected and a well-designed study can be carried out. In addition, it is very important to understand how to select and interpret diagnostic tests based on statistical methods.

REFERENCES

1. U.S. Department of Health and Human Services, Food and Drug Administration Center for Devices and Radiological Health. Policy for coronavirus disease-2019 tests during the public health emergency (revised) [Internet]. Silver Spring, MD: Center for Devices and Radiological Health; 2020 May [cited 2020 Nov 30]. Available from https://www.fda.gov/media/135659/download.
2. Toft N, Jørgensen E, Højsgaard S. Diagnosing diagnostic tests: evaluating the assumptions underlying the estimation of sensitivity and specificity in the absence of a gold standard. Prev Vet Med 2005;68:19-33.
PUBMED | CROSSREF

3. Garrett ES, Eaton WW, Zeger S. Methods for evaluating the performance of diagnostic tests in the absence of a gold standard: a latent class model approach. Stat Med 2002;21:1289-307.
PUBMED | CROSSREF

4. Florkowski CM. Sensitivity, specificity, receiver-operating characteristic (ROC) curves and likelihood ratios: communicating the performance of diagnostic tests. Clin Biochem Rev 2008;29 Suppl 1:S83-7.
PUBMED

5. Cardoso JR, Pereira LM, Iversen MD, Ramos AL. What is gold standard and what is ground truth? Dental Press J Orthod 2014;19:27-30.
PUBMED | CROSSREF

6. Pepe MS. The Statistical Evaluation of Medical Tests for Classification and Prediction. New York, NY: Oxford University; 2003.

7. Eusebi P. Diagnostic accuracy measures. Cerebrovasc Dis 2013;36:267-72.
PUBMED | CROSSREF

8. U.S. Food and Drug Administration. Guidance for Industry and FDA Staff: Statistical Guidance on Reporting Results from Studies Evaluating Diagnostic Tests. Silver Spring, MD: U.S. Food and Drug Administration; 2007.

9. Zhou XH, Obuchowski NA, McClish DK. Statistical Methods in Diagnostic Medicine. New York, NY: John Wiley; 2002.