Diagnostic accuracy – Part 1
Basic concepts: sensitivity and specificity, ROC analysis, STARD statement

June 2009

Ana-Maria Simundic
University Department of Chemistry
University Hospital SESTRE MILOSRDNICE
School of Medicine, Faculty of Pharmacy and Biochemistry,
Zagreb University
Vinogradska 29
10 000 Zagreb
CROATIA

The discriminative ability of a diagnostic procedure is called diagnostic accuracy, and a number of quantitative measures out of which sensitivity and specificity are mostly used in the biomedical literature can express it.

Each diagnostic-accuracy measure relates to some specific aspects of a diagnostic procedure. While some measures are used to assess the discriminative property of the test, others are used to assess its predictive ability.

Discriminative measures are mostly used by health-policy decision makers; predictive measures are most useful for predicting the probability of a disease in an individual.

Some measures assess the global performance of a test, whereas others are related to its ability to detect or exclude the disease, or to the clinical significance of a positive or negative test result in a specific patient.

Furthermore, measures of a test performance are not fixed indicators of a test quality, but are very sensitive to the characteristics of the population in which the test accuracy is being 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 outmost importance to understand the meaning of different measures of diagnostic accuracy and to know how to interpret them and under what conditions they may be used.

What is diagnostic accuracy

To discriminate the diseased from those who are healthy is the ultimate goal of every diagnostic procedure. What we would expect from an ideal biochemical marker is

that almost all healthy individuals shall have their values somewhere within the reference limits, whereas those who have a disease shall have significantly higher (less frequently lower) values of a measured parameter.

What we would expect to observe rather rarely are healthy individuals with an elevated marker concentration (the so-called false positives) as well as diseased individuals with values falling within the reference interval (false negatives).

Even though it may seem as an easy *"mission"*, the absolutely ideal marker does not exist and we therefore unfortunately always end up with a certain proportion of individuals having falsely elevated or lowered marker concentration.

The less of those false positives and false negatives observed, the better is the marker.

The only question is: how to measure this discriminative potential of some diagnostic procedure (biochemical parameter, panel of parameters, radiologic analysis or clinical exam)? How to know which procedure is better?

The discriminative ability of a diagnostic procedure is called diagnostic accuracy, and the number of quantitative measures out of which sensitivity and specificity are mostly used in the biomedical literature can express it.

Measures of diagnostic accuracy are:

- Sensitivity (Se)
- Specificity (Sp)
- Positive predictive value (PPV)
- Negative predictive value (NPV)
- Likelihood ratio (LR)
- Area under the ROC curve (AUC)
- Youden index
- Diagnostic odds ratio (DOR)

Why do we have so many measures of diagnostic accuracy

Each measure of diagnostic accuracy relates to some specific aspects of a diagnostic procedure. While some measures are used to assess the discriminative property of the test, others are used to assess its predictive ability.

Discriminative measures are mostly used by health-policy decision makers, whereas predictive measures are most useful for predicting the probability of a disease in an individual.

Some measures assess the global performance of a test, whereas others are related to its ability to detect or exclude the disease, or to the clinical significance of a positive or negative test result in a specific patient.

What is also important is the fact that measures of a test performance are not fixed indicators of a test quality. On the contrary, measures of diagnostic accuracy are very sensitive to the characteristics of the population in which the test accuracy is being 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 understand the meaning of different measures of diagnostic accuracy and to know how to interpret them and under what conditions they may be used.

How to assess the diagnostic accuracy of a biochemical marker

Let us imagine that we want to evaluate the diagnostic accuracy of S-100B, a new potential marker for acute ischemic stroke. How would you assess its diagnostic accuracy?

Measures of diagnostic accuracy are extremely sensitive to the design of the study aimed to assess the diagnostic accuracy of a certain marker.

Studies suffering from some major methodological shortcomings can severely over- or underestimate the indicators of test performance and limit the external validity of the study, i.e. the generalizability of the results of the study.

The easiest and most appealing way to design a diagnostic-accuracy study is a so-called *"two-gate"* (case-control) study design. In such studies, patients are compared with healthy individuals.

This way, measures of diagnostic accuracy have been shown to overestimate the measures severalfold, compared with properly designed studies that use single series of consecutive patients to evaluate the same test. The case-control study design is therefore not recommended.

In the properly designed study, patients are collected as a consecutive series of individuals in whom the target condition is suspected. The biochemical marker under evaluation is performed in all individuals presenting with disease symptoms.

Subsequently, the presence of disease is determined by performing the reference standard method for diagnosis.

In our example with a new marker (S-100B) for acute ischemic stroke, the ideal design would be as follows:

All individuals with acute ischemic stroke symptoms presenting to the Emergency department of our Neurology clinic are consecutively recruited into the study. Blood samples are drawn immediately and sent to the laboratory for S-100B concentration measurement.

All individuals undergo the same diagnostic work-up and a stroke diagnosis is made based on established criteria, equal for all patients.

Subsequently, statistical analysis is performed and measures estimated in order to assess the power of the S-100B marker to discriminate between individuals with and without acute ischemic stroke.

A collaborative group of researchers have developed the STARD (Standards for Reporting of Diagnostic Accuracy) statement aimed to improve the quality of reporting of studies of diagnostic accuracy.

The statement consists of a checklist of 25 items and a flow diagram that authors can use to ensure that all relevant information is present.

The aim and history of STARD as well as the STARD checklist, STARD flow diagram and many other related documents can be accessed at the official STARD website: stard-statement.org. The STARD initiative was a very important step toward the improvement of the quality of reporting of studies of diagnostic accuracy.

According to the STARD statement, the simple example of the flow diagram for our study of diagnostic accuracy of S-100B for acute ischemic stroke would be as presented on the FIGURE 1.

Calculating and interpreting sensitivity and specificity

A perfect diagnostic marker for acute ischemic stroke would have the potential to completely discriminate individuals with and without stroke. Unfortunately, as was already pointed out, such perfect diagnostic test does not exist.

Therefore, by using the cut-off for S-100B of 0.5 µg/L, for example, we may classify study participants into four subgroups considering parameter concentrations:

- True positive (TP) – subjects having stroke and S-100B > 0.5 µg/L
- False positive (FP) – subjects without stroke and S-100B > 0.5 µg/L
- True negative (TN) – subjects without stroke and S-100B < 0.5 µg/L
- False negative (FN) – subjects having stroke and S-100B < 0.5 µg/L

The first step in calculating sensitivity and specificity is to make a 2 × 2 table with groups of subjects divided

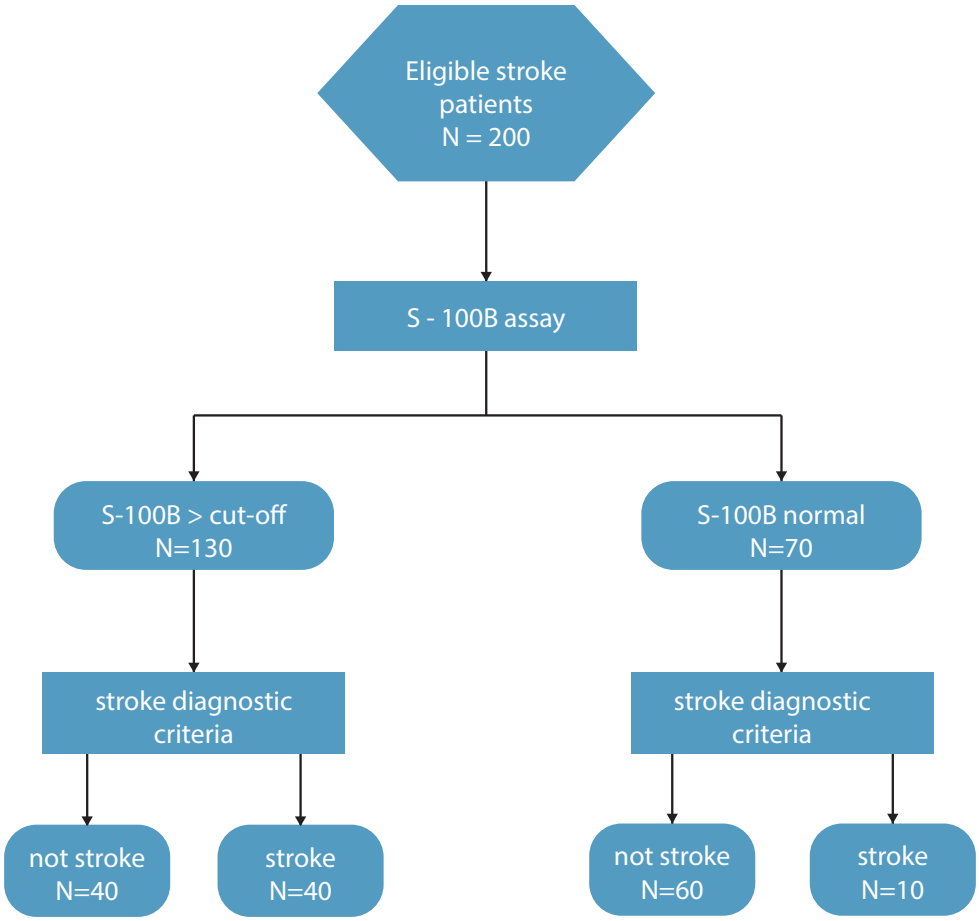

FIGURE 1: Flow diagram according to the STARD statement

according to a gold standard or reference method (diagnostic criteria) in columns, and categories according to test (S-100B) in rows (TABLE 1).

	Individuals with stroke	Individuals without stroke
S-100B > 0.5 µg/L	TP (N = 90)	FP (N = 40)
S-100B < 0.5 µg/L	FN (N = 10)	TN (N = 60)

TABLE 1: 2 × 2 table for calculating measures of diagnostic accuracy

Sensitivity (%) defines the proportion of true positive subjects with the disease in a total group of subjects with the disease (TP / (TP + FN)). In other words, 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 identify subjects with the disease.

In our example the sensitivity is 90 % at a cut-off value for serum S-100B protein of 0.5 µg/L.

What does it mean? It means that if we measure the S-100B concentration in every individual presenting with stroke symptoms at the Emergency department of our Neurology clinic, we shall observe S-100B > 0.5 µg/L in nine out of 10 individuals in whom stroke was subsequently diagnosed, according to standard diagnostic criteria for acute ischemic stroke (gold standard).

Moreover, it also means that if we solely rely on the S-100B result, in the absence of other diagnostic options, we would miss one out of every 10 stroke

patients. The question is: are we willing to accept such diagnostic uncertainty?

So, the sensitivity is a very useful marker that gives us an idea about the discriminative power of the marker and the proportion of diseased individuals missed by the marker.

However, what would be far more informative for the physician is: if a concentration of S-100B > 0.5 µg/L is measured in an individual presenting with stroke symptoms, how sure can I be that this patient has a stroke?

Unfortunately, sensitivity tells us nothing about it.

Specificity (%) is another measure of the diagnostic test accuracy, complementary to sensitivity. It is defined as a proportion of subjects without the disease with a negative test result in total of subjects without the disease (TN / (TN + FP)).

Analogous to sensitivity, specificity represents the probability of a negative test result in a subject without the disease.

Therefore, we can postulate that specificity relates to the aspect of diagnostic accuracy that describes the test ability to identify subjects without the disease, i.e. to exclude the condition of interest.

Again, let us look back at the example with stroke patients and the S-100B diagnostic marker. The specificity in our study turned out to be 60 % at a cut-off value for serum S-100B protein of 0.5 µg/L. What does it mean?

A specificity of 60 % means that if we measure the S-100B concentration in every individual presenting with stroke symptoms at the Emergency department of our Neurology clinic, in six out of 10 individuals in whom stroke was subsequently ruled out, a concentration of S-100B < 0.5 µg/L shall be observed.

It also means that four out of 10 individuals without stroke shall have a falsely elevated marker concentration. These individuals would be exposed to further diagnostic work-up and psychological stress related to the (spurious) existing probability of having a disease.

The question again is: are we willing to accept this diagnostic uncertainty? The answer is not an easy one, nor is there a unique answer to this question.

The decision on the acceptable level of diagnostic uncertainty depends on the disease characteristics, healthcare costs and psychological impact of a missed diagnosis and many other issues.

If a disease is a serious life-threatening condition, we may not want to miss it, so maximum sensitivity shall be most suitable.

So, the specificity also gives us an idea about the discriminative power of the marker. Again, as with sensitivity, what the physician would like to know is: if a concentration of S-100B < 0.5 µg/L is measured in an individual presenting with stroke symptoms, how sure can I be that this patient does not have a stroke?

The knowledge about the marker specificity does not provide the exact evidence for such clinical judgments.

ROC curves

The specificity and sensitivity of every diagnostic test depend on the selected cut-off level. Therefore, a pair of diagnostic sensitivity and specificity values exists for every individual cut-off. The ROC (Receiver Operating Characteristic) curve is constructed by plotting these pairs of values on the graph with the 1-specificity on the x-axis and sensitivity on the y-axis.

The shape of the ROC curve and the area under the curve (AUC) help us estimate the discriminative power of a test. The closer the curve follows the upper left-hand corner and the larger the area under the curve, the better the test is at discriminating between those with and without the disease.

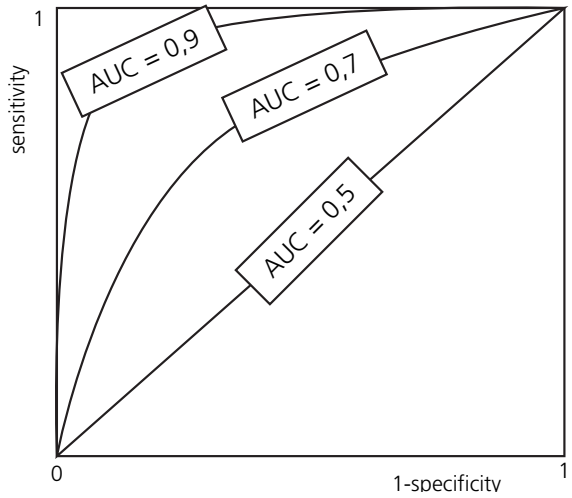

FIGURE 2: ROC curve

AUC is a global measure of diagnostic accuracy. The area under the curve may be any value between 0 and 1 and it is a good indicator of the overall quality of the test.

By comparing the areas under the two ROC curves we can estimate which test is better at diagnosing a disease. A perfect diagnostic test has an AUC of 1.0, whereas a useless test has an area ≤0.5. The interpretation of the AUC is described in TABLE 2.

AUC	Diagnostic accuracy
0.9-1.0	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

TABLE 2: The interpretation of the AUC curves

Conclusion

It is important to mention that neither sensitivity nor specificity is influenced by the disease prevalence, meaning that results from one study could easily be transferred to some other setting with a different prevalence of the disease in the population.

Nonetheless, sensitivity and specificity may vary greatly depending on the spectrum of the disease in the studied group. Sensitivity and specificity are commonly used estimates of diagnostic accuracy.

They should be well understood and carefully interpreted in order to serve as valid evidence for health care providers, clinicians and laboratory professionals; to the best for the patient care.

References

1. Irwig L, Bossuyt P, Glasziou P, Gatsonis C, Lijmer J. De signing studies to ensure that estimates of test accuracy are transferable. BMJ. 2002; 324(7338): 669-71.

2. Raslich MA. Markert RJ, Stutes SA. Selecting and interpreting diagnostic tests. Biochemia Medica 2007; 17(2): 139-270.

3. Rutjes AW, Reitsma JB, Di Nisio M, Smidt N, van Rijn JC, Bossuyt PM. Evidence of bias and variation in diagnostic accuracy studies. CMAJ. 2006; 14; 174(4): 469-76.

4. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, *et al*. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Clin Chem 2003; 49: 1-6.

5. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, *et al*. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaborati on. Clin Chem 2003; 49: 7-18.

6. Bossuyt PM. Clinical evaluation of medical tests: still a long road to go. Biochemia Medica 2006; 16(2) 89–228