# Biomarker assessment – what to be aware of

October 2012

**Suzanne Ekelund, MSc**
Principal Specialist Clinical Biochemist
Radiometer Medical ApS
Åkandevej 21
2700 Brønshøj , Denmark
Phone: +45 3827 3116
E-mail: suzanne.ekelund@radiometer.dk

Clinical biomarker studies seldom follow recommendations for the evaluation of new biomarkers.

Therefore it is important to assess any published clinical value of a biomarker or comparison of two or more biomarkers and see whether or not the published study applies to your situation. If the published study does not apply to your situation, you risk either:

- Accepting the use of a biomarker that does not give added clinical value, only added cost
- Rejecting the use of a biomarker that could give added clinical value

A few examples of limitations often seen are:

- Prevalence of disease in study population
- Low number of patients
- The type of patients used does not belong to the intended-use population
- The wrong cut-off is used
- Impact of comorbidities are not taken into account
- The interpretation of the study outcome is poorly proven
- No report of confidence intervals for performance measures

The limitations are often interwoven.

The effects of some of these limitations are shown.

## Glossary and abbreviations

| Abbreviation / term | Meaning |
|---|---|
| 95 % CI | 95 % confidence interval, i.e. the range within which the true value of a parameter is stated to lie with a 95 % probability |
| AUROC | Area under ROC curve, a way to reduce ROC performance to a single value representing expected performance; the value can be in the range 0-1 |
| Comorbidity | The existence of two or more diseases or conditions in the same individual at the same time |
| FN | False negative, number of sick persons with a negative test |
| FP | False positive, number of healthy persons with a positive test |
| NPV | Negative predictive value, fraction of test-negatives who do not have disease |
| PPV | Positive predictive value, fraction of test-positives who do have disease |
| Prevalence | Fraction of persons with disease in the tested population |
| ROC | Receiver operating characteristic, a plot of true positive fraction vs. false positive fraction for all potential cut-offs for a test |
| Sens | Sensitivity, i.e. fraction of persons with disease characterized as sick with the test in question |
| Spec | Specificity, i.e. fraction of persons without disease characterized as healthy with the test in question |
| Spectrum of disease | The range of disease states found in the patient population upon which the test is to be used |
| TN | True negative, number of healthy persons with a negative test |
| TP | True positive, number of sick persons with a positive test |

## Introduction

Many papers address the issue of how to evaluate a biomarker [1,2,3,4]. The recommendations for an objective and effective biomarker evaluation include [3]:

- Thorough analytical validation
- Evaluation of the evidence on associations between the biomarker and disease states
- Analysis of whether there is sufficient support for a specific use of the biomarker

It is also recommended that in later phases regulatory authorities ensure that expert panels reevaluate evidence.

These recommendations are for the ideal situation. However, in real life the recommendations are often not followed in studies. The reasons for this can be numerous. Eagerness to share new exciting data is one of them.

Therefore, when you use literature to assess the clinical value of a biomarker or to compare the clinical value to that of another biomarker, it is important to take into account whether or not the published study apply to your situation.

If the published study does not apply to your situation, you risk:

- Accepting the use of a biomarker that does not give added clinical value, only added cost
- Rejecting the use of a biomarker that could give added clinical value

## Evaluating clinical value of a parameter

Evaluating the usefulness of a parameter based on publications should be done with caution. The limitations of the published studies should be considered. A few examples of limitations are:

- Prevalence of disease in study population
- Low number of patients
- The type of patients used does not belong to the intended-use population
- The wrong cut-off is used
- Impact of comorbidities are not taken into account
- The interpretation of the study outcome is poorly proven
- No report of confidence intervals for performance measures

The limitations are often interwoven like when you have comorbidities in some patients and not in others, then you should perhaps have used a different cut-off for the patients with comorbidities, and in the end the interpretation of the study will turn out to be poorly proven.

## Choice of cut-off

The cut-off determines the clinical sensitivity (fraction of true positives to all with disease) and specificity (fraction of true negatives to all without disease). The choice of cut-off is always a trade-off between sensitivity and specificity.

This can be seen in Fig. 1 and Fig. 2. When 400 µg/L is chosen as the analyte concentration cut-off, the sensitivity is 100 % and the specificity is 54 %. When the cut-off is increased to 500 µg/L, the sensitivity decreases to 92 % and the specificity increases to 79 %.

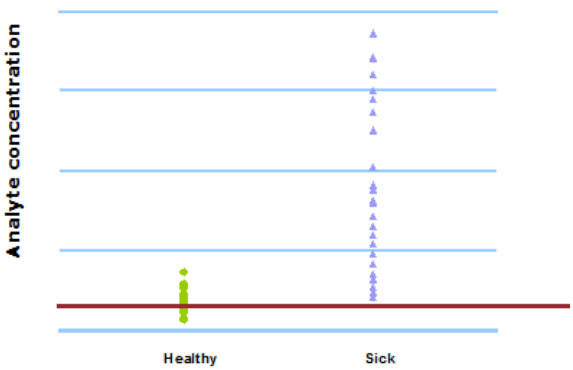


FIG. 1: Cut-off = 400 µg/L

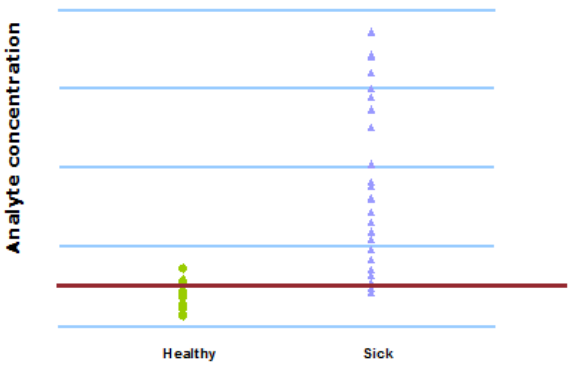


FIG. 2: Cut-off = 500 µg/L

However, sometimes you have to apply different cut-offs to different patient groups. An example: we want to use a new parameter to differentiate patients who are infected from those who are not infected.

We try to find a suitable cut-off but as can be seen in Fig. 3, there is no ideal cut-off due to overlap of the ranges of results from the infected and the non-infected group.

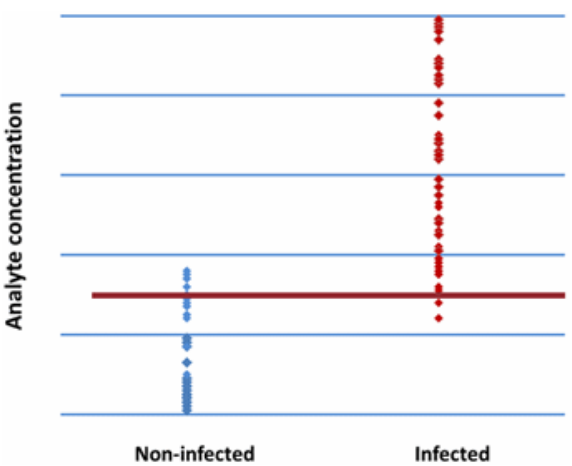


FIG. 3: Same cut-off for all patients

But in this case the parameter tested is not only an infection marker but also an acute-phase reactant. This means that surgery will increase the level, also without an infection.

Let us try instead to divide the patients into medical and surgical subgroups and further divide these subgroups into infected and non-infected.

Now, as can be seen in Fig. 4, we are able to define separate cut-offs for the medical and the surgical groups, respectively. Now the parameter seems to be useful in the differentiation of infected from non-infected patients.

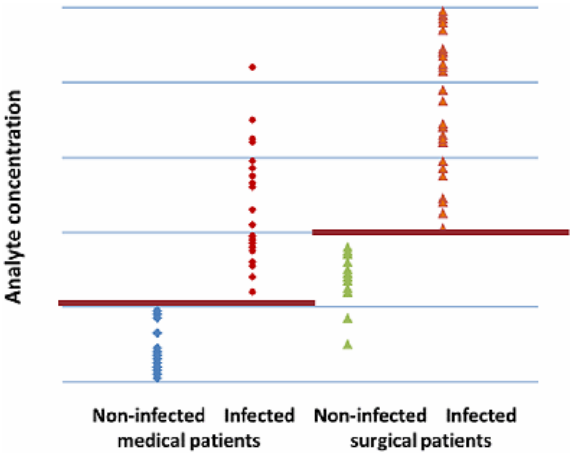


FIG. 4: Patient-group-specific cut-offs

It should always be kept in mind that a cut-off has to be chosen based on one population and tested on another population. A reason for this is that a data-driven choice of cut-off often boosts the diagnostic performance of a biomarker.

## Study population

The ideal biomarker comparison is based on results obtained with both/all biomarkers on the same samples [5].

When separate studies are compared, there can be biases which have affected the selection of persons for the different groups. Thus apparent differences in biomarker performance may instead reflect differences between the groups tested.

The spectrum of disease could be different in the groups compared. One group might have more persons with advanced disease, which might be more easily detected. Another group might have a majority of persons with minimal disease, which might be harder to detect.

However, it should also be kept in mind that even though a comparison based on results obtained with both/all biomarkers on the same samples can be directly used to compare the relative performances of the biomarkers, the relative performances may be different if the biomarkers are tested on another population with a different prevalence, a different spectrum of disease, etc.

It is often seen that a biomarker is tested in healthy persons and in persons with a specific disease. If the values obtained in the two groups differ significantly, it is concluded that the biomarker could be useful as an aid in the diagnosis of that specific disease.

However, this is not necessarily the case. The biomarker may be useful to differentiate persons with the specific disease from healthy controls, but persons with the same symptoms as those with the specific disease but with another diagnosis might have the same levels of the biomarker as those with specific disease. In such a case the interpretation of the study outcome is poorly proven.

Comparing biomarker values in healthy and in persons with a specific disease can be seen as a pilot study to show if the biomarker may have some potential in a specific disease. However, to claim usefulness in diagnosis the study population has to be the intended-use population where there is a need to differentiate among the patients.

## Predictive values

The predictive value of a test is a measure in percentage of the times that the value (positive or negative) is the true value, i.e. the percent of all with a positive test who are actually sick is the positive predictive value (PPV) and the percent of all with a negative tests who are not sick is the negative predictive value (NPV).

| | | Disease (the "truth") | | Predictive values |
|---|---|---|---|---|
| | | + | - | |
| Test | + | True positives (TP) | False positives (FP) | PPV = TP / (TP + FP) |
| | - | False negatives (FN) | True negatives (TN) | NPV = TN / (TN + FN) |
| | | Sensitivity = TP / (TP + FN) | Specificity = TN / (TN + FP) | Prevalence= (TP + FN) / (TP + FN + TN + FP) |

TABLE I: Comparing a method with the "truth"

As can be seen in Table I the sensitivity, which determines the number of true positives, and the specificity, which determines the number of true negatives, both have a strong impact on the predictive values.

However, the prevalence, i.e. the fraction of sick persons in the population tested, also has a strong impact on the predictive values of a test.

## Influence of prevalence on predictive values

As can be seen in Fig. 5 PPV increases with increasing disease prevalence, whereas NPV decreases with increasing disease prevalence. To understand why that is, consider a situation where everyone in a population is sick (prevalence = 100 %).

In this situation, every positive result would be a true positive and there are no false positive results. Then the PPV would be 100 %. Conversely, if no one in the population is sick (prevalence = 0 %), every positive result would be a false positive.

As there are no true positives, the PPV is 0 %. Thus we can see that the disease prevalence influences the PPV by influencing the true positive and false positive rates.

Similar arguments can be used for NPV. When there are only sick persons in a population, every negative result would be false negative, and if there are only healthy persons in a population, every negative result would be true negative.

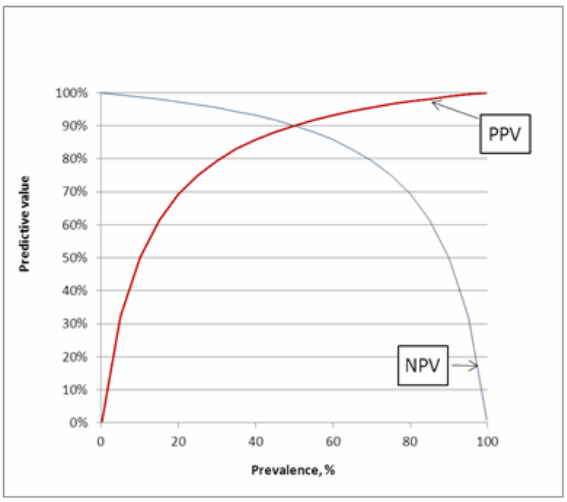


FIG. 5: Predictive values as function of prevalence.
Assay sensitivity = 90 % and assay specificity = 90 %

So what do predictive values versus prevalence tell us?

We can choose prevalence; say 10 %. It means that with 100 persons in the study, we have 10 who are sick and 90 who are healthy.

If the assay has sensitivity 90 %, then 9 of the sick persons (90 % of 10 persons) will get a true positive test and one (10 – 9) will get a false negative test. If the assay has specificity 90 %, then 81 of the healthy persons (90 % of 90 persons) will get a true negative test and 9 will get a false positive test. The results are illustrated in Fig. 6.

FIG. 6: Test results Results of sick persons are colored red, results of healthy persons are colored blue. Positive test result shown as +, negative test result as ˙.

When we look at Fig. 6, we can see who is sick and who is healthy because they have different colors. However, when we get a test result, we are "color blind" and only see whether it is positive or negative like in Fig. 7.

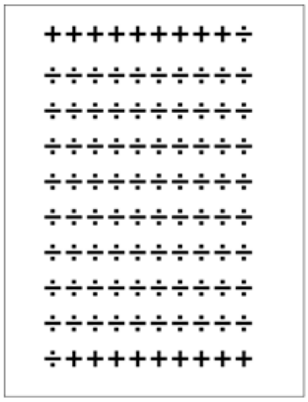


FIG. 7: Test results Positive test result shown as +, negative test result as ÷.

We have 18 + and 82 ÷. From the calculations above we know that only 9 of the 18 + are sick and only 81 of the 82 ÷ are healthy. From this information we can calculate the predictive values:

PPV = 100 % × 9/18 = 50 %

NPV = 100 % × 81/82 = 99 %

A positive predictive value of 50 % means that in half of the cases where we get a positive result the person is sick and in half of the cases the person is healthy. Similarly, a negative predictive value of 99 % means that when we get a negative result, the patient is in 99 % of the cases healthy and only in 1 % of the cases the person is sick.

These predictive values we can easily find directly when we use Fig. 8.

So what Fig. 5 and Fig. 8 show us is that it is important to remember that the predictive values are not universal but depend on each specific clinical setting.

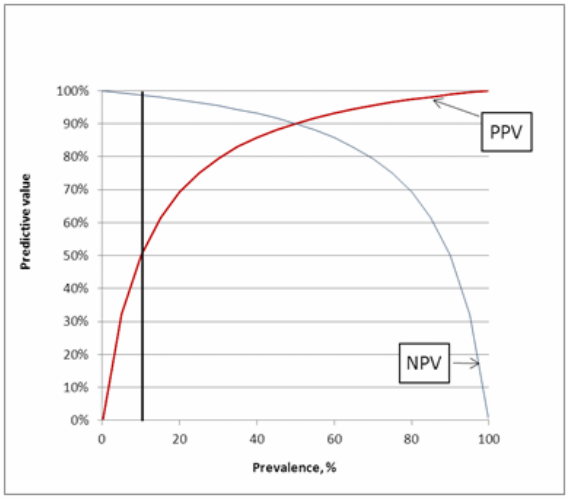


FIG. 8: Predictive values as function of prevalence

## Influence of sensitivity and specificity on predictive values

The influence of sensitivity can be seen in Fig. 9. Lowering the sensitivity from 90 % to 60 % has an impact on the predictive values. The greatest impact is on the negative predictive value because lowering the sensitivity means increasing the number of false negatives.

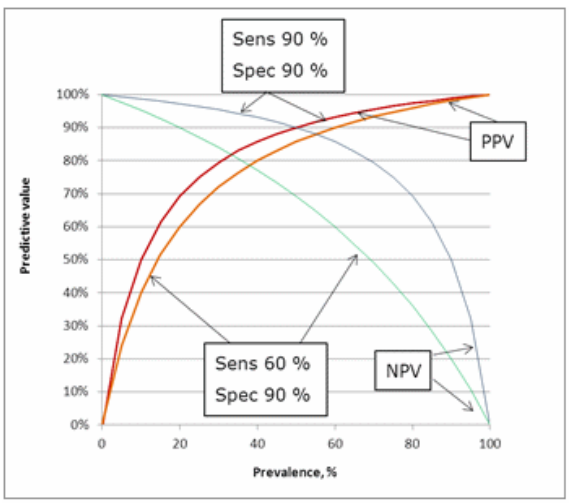


FIG.9: Influence of sensitivity on the predictive values

The influence of specificity can be seen in Fig. 10. Lowering the specificity from 90 % to 60 % has an impact on the predictive values. The greatest impact is on the positive predictive value because lowering the specificity means increasing the number of false positives.

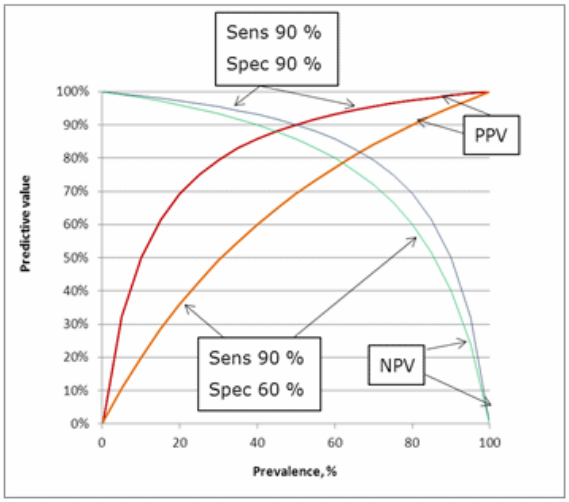

FIG. 10: Influence of specificity on the predictive values

## Influence of total number of patients in a study

The CLSI EP12-A2 guideline [6] recommends that as a minimum testing should continue until results from at least 50 positive specimens are obtained and that at least 50 negative specimens using the comparative method should be obtained to determine the specificity of the candidate method.

Many published studies have fewer numbers of either positive and/or negative specimens. The width of the confidence interval gives us an idea about how uncertain we are about a parameter and an advantage of a high number of participants in a study is that the confidence interval for the performance measures gets narrower with an increasing number of participants.

We can make a table like Table I where a method is compared to the "truth" and calculate the sensitivity and the specificity with the 95 % confidence intervals and use it to see an example of the influence of the number of patients included in a study.

It can be seen in Table II and Table III that when the number of patients is increased (the fractions of TP, TN, FP and FN in the example stay the same), the confidence intervals get narrower and thus the actual information obtained from the study gets more valuable.

In this example the sensitivity goes from a spread of 20.1 percentage points to 5.8 percentage points when the number of participants is increased 10-fold. Because the example has an amount of healthy persons that is more than three times as high as the amount of sick persons, the confidence interval for the specificity is narrower and the change when the number of participants is increased 10-fold only goes from 5.9 percentage points to 1.7 percentage points.

The confidence intervals for the predictive values will narrow similarly.

| n=145 | | Disease (the "truth") | | Sensitivity, % (95 % CI) | Specificity, % (95 % CI) |
|---|---|---|---|---|---|
| | | + | - | | |
| Test | + | 30 | 12 | 90.9 (77.4 – 97.5) | 89.3 (85.3 – 91.2) |
| | - | 3 | 100 | | |

TABLE II: Sensitivity and specificity confidence intervals, n = 145

| n = 1450 | | Disease (the "truth") | | Sensitivity, % (95 % CI) | Specificity, % (95 % CI) |
|---|---|---|---|---|---|
| | | + | + | | |
| Test | + | 300 | 120 | 90.9 (87.4 – 93.5) | 89.3 (88.3 – 90.0) |
| | - | 30 | 1000 | | |

TABLE III: Sensitivity and specificity confidence intervals, n = 1450

## Area under roc curve

The CLSI EP812-A2 guideline [6] mentions that an appropriate measure to describe diagnostic accuracy includes ROC curves.

An ROC curve shows the relationship between clinical sensitivity and specificity for every possible cut-off. The ROC curve is a graph with:

- The x-axis showing 1 – specificity (= false positive fraction = FP/(FP+TN))
- The y-axis showing sensitivity (= true positive fraction = TP/(TP+FN))

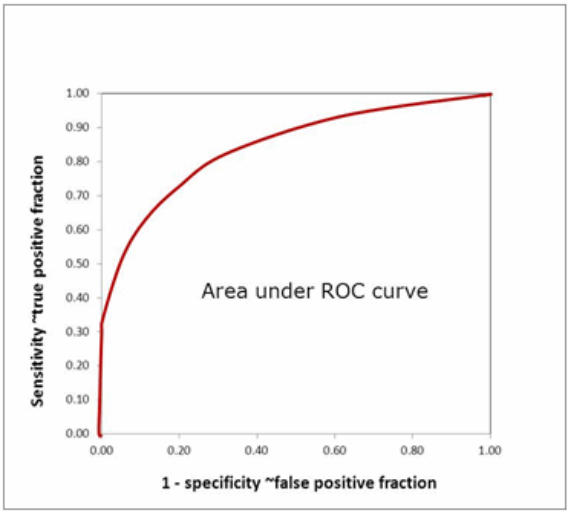


FIG. 11: Area under ROC curve

Thus every point on the ROC curve represents a chosen cut-off even though you cannot see this cut-off on the graph. What you can see is the true positive fraction and the false positive fraction that you will get when you choose this cut-off.

The area under the ROC curve (AUROC) of a test is often used as a criterion to measure the test's discriminative ability, i.e. how good is the test in a given clinical situation. The closer an ROC curve is to the upper left corner, the more efficient is the test and the closer to 1.0 the AUROC will be.

AUROC is also often used in comparisons between biomarkers. The advantage of using AUROC to compare biomarkers is that the AUROC reveals real differences in diagnostic performance and not just differences in sensitivity and specificity due to different choices of cut-off.

When comparing AUROCs, confidence intervals need to be taken into consideration. The confidence interval of the difference between two AUROCs is determined by numbers of healthy and sick participants in the study and whether or not the two AUROCs were obtained on the same specimens or different specimens. If a confidence interval for the difference between AUROCs contains the value 0.0, then the two AUROCs do not differ significantly.

In Table IV we can see that if we have a small study with 50 healthy persons and 20 sick persons and we compare two biomarkers using the same specimens, then biomarker 1 with AUROC 0.90 and biomarker 2 with AUROC 0.81 cannot be claimed to be significantly different from a diagnostic point of view (under the specified circumstances of the study). If biomarker 2 had had AUROC 0.80, they would have been claimed significantly different.

| | AUROC 1 | AUROC 2 | Difference (95 % CI) |
|---|---|---|---|
| Same specimens | 0.90 | 0.81 | 0.09 (-0.005 – 0.185) |
| Different studies/ specimens | 0.90 | 0.74 | 0.16 (-0.007 – 0.327) |

TABLE IV: Comparing AUROC; 50 healthy, 20 sick in all studies

In Table IV we can also see that if we have two small studies, each having 50 healthy persons and 20 sick persons and we compare two biomarkers, then biomarker 1 (study 1) with AUROC 0.90 and biomarker 2 (study 2) with AUROC 0.74 cannot be claimed to be significantly different from a diagnostic point of view (under the specified circumstances of the studies).

If biomarker 2 had had AUROC 0.73, they would have been claimed significantly different. In addition it can also always be discussed whether the two studies had groups with a similar spectrum of disease.

Thus it can be seen that when we have two separate studies, the difference between two areas under ROC curves will be greater before it is considered significant compared to the situation where we look at a difference between two ROC curves based on the same samples in the same study.

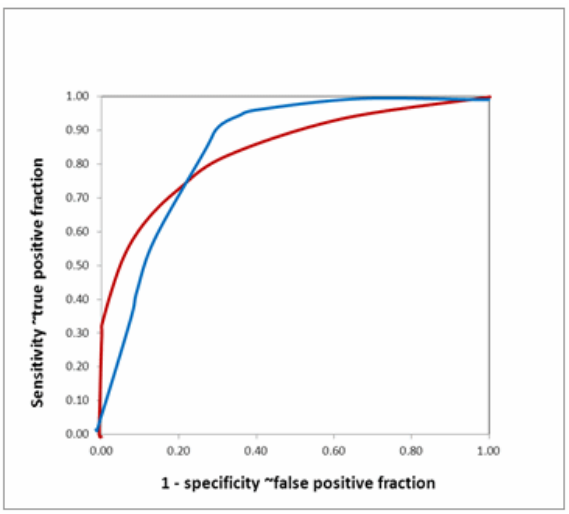


FIG. 12: Comparing area under ROC curve (hypothetical)

Equal AUROCs for two biomarkers show that their overall performances are similar. However, it does not mean that both the curves are identical. They may cross each other as in Fig. 12. The two ROC curves have nearly identical AUROC.

However, they cross each other. The biomarker with the red curve performs better than the biomarker with the blue curve when high specificity (low false positive fraction) is needed, while the biomarker with the blue curve performs better than the biomarker with the red curve when high sensitivity is needed.

It is justified to look at partial areas under ROC curves in such situations where AUROCs are nearly identical but differ in their diagnostic efficacy in specific segments.

If the biomarkers you are comparing are markers of a disease with a grave prognosis, you would want to have high specificity. Thus you would choose the biomarker with the red curve.

However, if the biomarkers you are comparing are used to select patients who need to undergo a confirmatory test for a disease that is important to treat, then you would choose the biomarker with the blue curve to have high sensitivity.

ROC curves with different total AUROCs may be similar in specific regions. In such cases it is also an advantage to look at partial AUROC.

## The influence on diagnostic performance

We now know that:

- Cut-off has an influence on diagnostic performance
- Cut-offs can be dependent on subgroups (comorbidities)
- Spectrum of disease has an influence
- Predictive values are highly dependent on
  - Prevalence
  - Sensitivity
  - Specificity
- Number of participants in a study has a great influence on the confidence interval for diagnostic performance measures
- When differences between AUROCs are compared, confidence intervals are important
- Partial AUROC analysis should be used when relevant
- To claim usefulness of a biomarker in diagnosis the study population has to be the intended-use population

## References


1. Pletcher MJ, Pignone M: Evaluating the clinical utility of a biomarker: a review of methods for estimating health impact. Circulation 2011; 123: 1116-24.

2. Dancey JE, Dobbin KK, Groshen S *et al*. Guidelines for the development and incorporation of biomarker studies in early clinical trials of novel agents. Clin Cancer Res 2010; 16(6): 1745-55.

3. Committee on Qualifications of Biomarkers and Surrogate Endpoints in Chronic Disease, Institute of Medicine. Evaluation of Biomarkers and Surrogate Endpoints in Chronic Disease. Washington, DC: The National Academies Press. 2010.

4. Ray P, Manach YL, Riou B *et al*. Statistical evaluation of a biomarker. Anesthesiology 2010; 112: 1023-40.

5. CLSI document EP24-A2 Assessment of the diagnostic accuracy of laboratory tests using receiver operating characteristics curves; approved guideline. – 2nd edition.

6. CLSI/NCCLS document EP12-A2 User protocol for evaluation of qualitative test performance; approved guideline - 2nd edition. Vol. 28 No. 3. 2008