Clinical Chemistry 58:9
1292–1301 (2012)

Review

Quantifying the Accuracy of a Diagnostic Test or Marker

Kristian Linnet,[1*] Patrick M.M. Bossuyt,[2] Karel G.M. Moons,[3] and Johannes B. Reitsma[3]

BACKGROUND: In recent years, increasing focus has been directed to the methodology for evaluating (new) tests or biomarkers. A key step in the evaluation of a diagnostic test is the investigation into its accuracy.

CONTENT: We reviewed the literature on how to assess the accuracy of diagnostic tests. Accuracy refers to the amount of agreement between the results of the test under evaluation (index test) and the results of a reference standard or test. The generally recommended approach is to use a prospective cohort design in patients who are suspected of having the disease of interest, in which each individual undergoes the index and same reference standard tests. This approach presents several challenges, including the problems that can arise with the verification of the index test results by the preferred reference standard test, the choice of cutoff value in case of a continuous index test result, and the determination of how to translate accuracy results to recommendations for clinical use. This first in a series of 4 reports presents an overview of the designs of single-test accuracy studies and the concepts of specificity, sensitivity, posterior probabilities (i.e., predictive values) for the presence of target disease, ROC curves, and likelihood ratios, all illustrated with empirical data from a study on the diagnosis of suspected deep venous thrombosis. Limitations of the concept of the diagnostic accuracy for a single test are also highlighted.

CONCLUSIONS: The prospective cohort design in patients suspected of having the disease of interest is the optimal approach to estimate the accuracy of a diagnostic test. However, the accuracy of a diagnostic index test is not constant but varies across different clinical contexts, disease spectrums, and even patient subgroups.

Diagnostic tests, like any other medical intervention, require proper evaluation before their introduction and recommendation for use in clinical practice. This is the first in a series of 4 reports that review various aspects of the test evaluation process. The major categories of test evaluation studies are presented in a schematic form in Fig. 1. The focus in this series will be on studies that examine the clinical validity and clinical utility of a test. The important step of examining the analytical or technical aspects and validity of a new test will not be discussed, but relevant information on these types of studies has been described elsewhere *(1)*.

We begin this series by considering the basic principles and the state of the art in the evaluation of the clinical accuracy of diagnostic biomarkers and other medical tests. In such accuracy studies the results of 1 or more tests under evaluation (i.e., index tests) are compared with the results of the prevailing clinical reference standard or method. This reference standard or method is the test or strategy that is clinically used to determine the presence or absence of the disease of interest (i.e., target disease). Such accuracy studies provide information about the degree of agreement in results from the index tests and the presence or absence of disease, i.e., the reference standard results. Additionally, these studies provide information about the frequency of types of errors [i.e., false-positive (FP)[4] and false-negative (FN) test results] by the index test compared to the reference standard.

Although an index test may provide a more accurate, more timely, or less invasive identification of the target disease compared to the reference standard, this does not automatically translate to improved therapeutic management, let alone to improved patient health or cost-effectiveness of care in general. Consequently, increasing attention is being paid to studies that examine the utility of a test in terms of these latter aspects or outcomes. These so-called clinical utility studies (Fig. 1) are directed at documenting the degree to which the actual use of a test leads to improved therapeutic management and consequently improved patient outcomes or, more generally, cost-effectiveness of provided care. The theory and conduct of such studies

[1] Section of Forensic Chemistry, Department of Forensic Medicine, Faculty of Health Sciences, University of Copenhagen, Copenhagen, Denmark; [2] Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Center, University of Amsterdam, Amsterdam, the Netherlands; [3] Julius Center for Health Sciences and Primary Care, UMC Utrecht, Utrecht, the Netherlands.

* Address correspondence to this author at: Section of Forensic Chemistry, Dept. of Forensic Medicine, Frederik Femte Vej 11, DK-2100 Copenhagen, Denmark. Fax +45-35326085; e-mail kristian.linnet@forensic.ku.dk.

Received January 17, 2012; accepted June 28, 2012.
Previously published online at DOI: 10.1373/clinchem.2012.182543

[4] Nonstandard abbreviations: FP, false positive; FN, false negative; DVT, deep venous thrombosis; TP, true positives; TN, true negatives; P(D), prevalence of disease; LR, likelihood ratio.

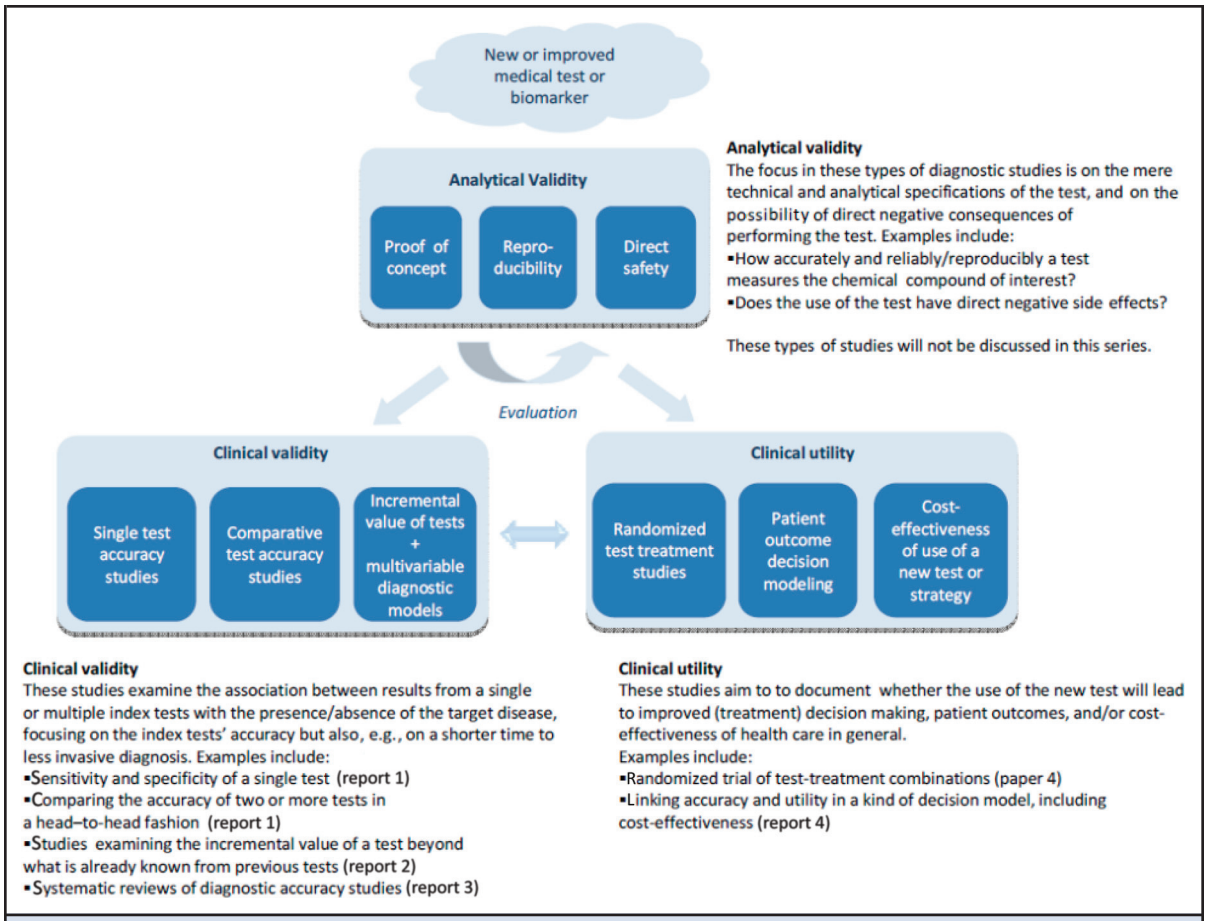

Fig. 1. Overview of the major classes of diagnostic test evaluations.

will be discussed in much more depth in a subsequent article.

Our aim in this first report is to discuss key issues in the design, conduct, analyses, and interpretation of results from studies that examine the diagnostic accuracy of a single test or compare the accuracy of 2 or more tests in a head-to-head fashion. In the second report we will discuss how to examine and quantify the added value of a new test beyond what is already known from previous tests. In a third report we will describe and discuss the main steps in performing a systematic review and metaanalyses of several primary studies examining the diagnostic accuracy of tests. The series will end with a report that provides a discussion of studies examining the clinical utility of a test.

The improved understanding of human metabolism and unravelling of the human genome has produced an avalanche of novel markers of disease and disease processes. We hope that this series will help the readers of *Clinical Chemistry* to better understand reports about the clinical evaluation of novel or existing diagnostic biomarkers and other laboratory tests.

Evaluation of a Single Test

The assessment of new tests, or reevaluation of commonly used tests, is an important area *(2, 3)*. Many tests are introduced to replace existing tests, and a systematic and unbiased approach for evaluation and comparison is important. We will consider first single-test accuracy studies and head-to-head comparisons in which the focus is on study design, type of accuracy measures, and data analysis. The core design of a diagnostic accuracy study is one in which a test under evaluation (i.e., the index test) is compared with a reference standard by applying both on the same individuals who are suspected of having the target disease of interest. In the data analysis, estimates of diagnostic performance are obtained and statistical computations are carried out to assist in the interpretation. The simplest situation is a comparison of a single index test to a reference standard (i.e., single-test accuracy study). A head-to-head comparison of 2 index tests with a common reference test (paired or randomized accuracy study) brings additional challenges, but also provides direct

information about differences in accuracy between the 2 index tests.

Empirical Example: Diagnosis of Suspected Deep Venous Thrombosis

Throughout this report, we illustrate concepts and methods using some of the empirical data from a previously published study in primary care patients suspected of having deep venous thrombosis (DVT), the target disease *(4, 5)*. In brief, the study included 2086 patients suspected of DVT, where DVT was defined as having at least 1 of the following symptoms: presence of swelling, redness, and/or pain in the leg. All patients underwent a standardized diagnostic work-up including medical history, physical examination, and testing for D-dimer, the index test. The reference standard or method consisted of repeated compression ultrasonography tests and was performed in all patients, independent of the results of the index test and blinded to these index test results. In total, 416 of the 2086 included patients (20%) had DVT. We note that for this report the data are used for illustration purposes only, and by no means to quantify the true diagnostic accuracy of the index test for the clinical problem at hand.

Design of Diagnostic Accuracy Studies

THE PROSPECTIVE COHORT DESIGN IN SUSPECTED PATIENTS

When studying the accuracy of a diagnostic test or marker to discriminate between the presence or absence of a particular disease in individuals suspected of having that target disease, it is commonly understood that the technical and analytical development of the test or assay has been completed, including method evaluations with assessments of, e.g., the measurement range, linearity, precision, and direct safety of the assay (Fig. 1, top). Further evaluation of the test is carried out in the clinically relevant context, first and foremost in the population for which it is intended to be used. Here, it is generally optimal to use the prospective cohort approach in suspected patients, in which the index is applied in a consecutive series of individuals suspected of having the target disease, on the basis presenting symptoms or signs. Subsequently, in each patient the reference standard is applied irrespective (i.e., independent) of the results of the index test, and also blinded to the results of the index test *(2, 6)*. In our example, when evaluating a test for the presence or absence of DVT (the target disease), a consecutive series of individuals seeking a general practitioner because of a red, swollen, and/or painful leg forms the intended patient population. A D-dimer test that can be carried out locally in general practice is the index test,

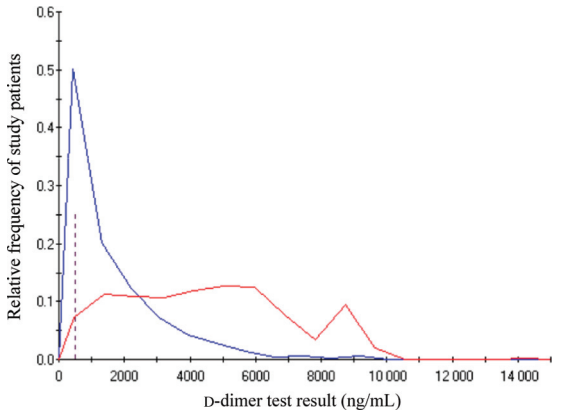

Fig. 2. Distribution of the quantitative D-dimer values for DVT and non-DVT participants from our example study.

Blue line, non-DVT; red line, DVT. The dashed line indicates the frequently used cutoff value of 500 μg/L.

and the reference test is the repeated radiological or ultrasonographic test carried out after referral to a hospital, in all study patients irrespective of and blinded to the results of the D-dimer index test. The interest in such clinical studies is how accurately the D-dimer assay can distinguish between individuals with and without DVT as classified by the prevailing, usually more burdensome or costly reference test.

For an ideal index test, there should be complete separation between index test values in the 2 groups, i.e., with and without the target disease. However, such separation is seldom observed. Usually, there will be some overlap in index test results suggesting that the index test is not perfect and cannot completely replace the reference test. In our example data, Fig. 2 shows how D-dimer values are distributed in those with (n = 416) and without (n = 1670) venous thrombosis as established by the prevailing reference test.

In addition to application of the reference standard in all study patients independent of and blinded for the results of the index test, it is also essential in this design that there is no nonrandom preselection of study participants. Participants are selected purely on the presence of predefined symptoms and/or signs of the target disease. The nondiseased group is made up of individuals initially suspected of having the target disease given these predefined symptoms/signs, who eventually were found not to have the target disease. Of course, they may have another underlying disease. Individuals in the diseased group, on the other hand, have the disease in the phase that naturally is presented to the doctor. In the context of diagnostic tests applied

by general practitioners, it will usually be in the early phase of the disease.

The result of a test evaluation is highly dependent on the clinical setting. One assessment may be obtained in general practice, but another in a hospital with patients who have a more advanced degree of disease. The importance of the spectrum of disease was pointed out by Ransohoff and Feinstein *(7)*. Numerous promising diagnostic tests have been introduced and evaluated with very good results on patients with advanced disease. Later on, when the tests have been applied on patients with early disease, the results have been disappointing. This holds, for example, for many tumor markers which, after showing promising results in initial studies of advanced disease have generally turned out to be worthless in screening individuals for possible preclinical disease. The prevalence of disease in a study may give a hint concerning the spectrum of disease. For a disease with a low prevalence, it is likely that patients with early disease are included, whereas studies in which the prevalence of disease is very high, e.g., larger than 0.5, are likely to be dominated by patients with advanced disease *(8*–10*)*. Thus, it is preferable that the stage of disease in the studied patients be reported. As we will discuss and illustrate below, the accuracy of a diagnostic index test is not constant but often differs across clinical contexts, disease spectra, and indeed patient subgroups *(8, 9, 11)*.

Also, in an evaluation of a diagnostic test, it is critical that all study participants are correctly classified as diseased or nondiseased. For example, in case of a tumor marker index test, the reference standard is commonly a histological diagnosis. If all participants get a histologically confirmed diagnosis, independent of and blinded for the index marker results, there will be no problems. However, if only individuals having a positive tumor marker index test result are subjected to a biopsy, we do not have the same reference standard result for all participants, a situation known as selective partial or differential disease verification *(12, 13)*. In this case, only the specificity of the test can be evaluated, because only the number of FP are estimated, not the number of FN. Partial or differential verification is likely in a screening situation, in which it is often unethical or even impossible to refer all participants for reference testing. In other diagnostic areas, one should thus use the same and best available diagnostic procedure as reference standard, in all study participants, independent of and blinded for the results of the index test.

THE CASE CONTROL DESIGN

Although the prospective suspected patient cohort design generally is regarded as the optimal approach for evaluating the accuracy of a diagnostic test, the design may not be practical in all cases. If the disease in question is very rare, a very large nondiseased group is included with a small diseased group. A more economic and practical approach in this circumstance is to compare test values in a diseased group with a control group of similar size or perhaps double or triple the size. This design is equivalent to the case control design frequently used in epidemiology *(14, 15)*. The same points concerning selection mentioned above are of importance here. How the diseased group was selected, e.g., hospitalized or outpatients, and how the control group was obtained should be clearly indicated. Ideally, the controls are a random sample of the same underlying suspected patient population (based on the same predefined symptoms and signs) that the cases came from, a so-called nested case control approach in epidemiology *(14, 15)*. Use of a control population that does not fit this criterion, but rather is a sample of healthy persons or individuals with some other disease, may lead to biased or clinically unrepresentative estimates of the accuracy measures of the index test *(14*–*16)*.

Disease status		
Test result	Diseased	Nondiseased
Positive	TP	FP
Negative	FN	TN

Fig. 3. The basic 2-by-2 table for estimating the diagnostic accuracy of a dichotomous or dichotomized quantitative test result.

Positive test results are divided into TP and FP, and negative results into TN and FN.

Measures of Diagnostic Accuracy

TWO TYPES OF ERROR

The ideal diagnostic test classifies all individuals correctly as diseased or nondiseased, and the error rate is zero. This is the case when there is no overlap between test values in the 2 groups. However, when test values in the nondiseased and diseased groups overlap, some individuals are likely to be misclassified by the test (Fig. 2). To use a quantitative test to classify individuals as diseased or nondiseased some suitable cutoff value should be selected. Results of an index test that exceed the cutoff in individuals in whom the disease is truly present (as independently and blindly confirmed by the reference standard) are defined as true positives (TP) (Fig. 3). Similarly, index test values below the cutoff in truly nondiseased individuals are true negatives (TN). Correspondingly, index test values below the

cutoff in truly diseased individuals are FN, and index test values above the cutoff in truly nondiseased individuals are FP. An overall error rate or nonerror rate can then be assessed. The overall accuracy of an index test can then be defined as the proportion of true classifications out of all classifications:

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

This overall nonerror rate can be subdivided into the nonerror rate of the nondiseased individuals, which is called the specificity of the test:

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

A very specific test provides negative results for all or almost all individuals free of the target disease.

The nonerror rate of the diseased group is the sensitivity of the test:

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

A sensitive test detects all or almost all diseased individuals.

Given our example study concerning the D-dimer assay with a frequently used cutoff of ≥500 μg/L (dashed line, Fig. 2), the sensitivity was 0.97 (3% of the individuals with DVT had a value <500 μg/L), and the specificity was 0.37. The overall accuracy was 0.50. Thus, this test is rather sensitive, detecting all but 3% of those having DVT. On the other hand, the specificity is rather low, resulting in many FP results.

To assess the (im)precision of these accuracy estimates, either CIs for the estimates or SEs are needed. If the cutoff value of a quantitative index test is fixed and not dependent on the results obtained in the study, simple statistical procedures based on the binomial distribution can be applied. Given random sampling, the 95% CI of a proportion can be obtained from tables or computer programs. Quite often an approximation of the binomial to the normal distribution is used for estimation of the 95% CI of proportions, as ±2 SE(P), where $\text{SE}(P) = [P(1 - P)/N]^{0.5}$ (P, proportion; N, sample size).

Unfortunately, the normal approximation does not work well in small samples or when proportions are close to 0 or 1. Both situations occur regularly in diagnostic research. The method of Wilson is a good alternative *(17)*. Table 1 shows the widths of the 95% CIs at various sample sizes of 20–1000 for 2 selected proportions, either a sensitivity or a specificity of an index test. For example, at a sample size of 20, the 95% CI ranges from 0.56 to 0.94 for a proportion of 0.80. Thus, at small sample sizes, only rather uncertain estimates of specificity or sensitivity are obtained. Bachmann et al. *(18)* reported that for 43 nonscreening studies on diagnostic accuracy of tests, the median sample size was 118 (interquartile range 71–350). The median for the diseased group was 49 (interquartile range 28–91) and for the nondiseased 76 (interquartile range 27–209). Concerning our D-dimer example study, the sample size was rather high, and so the estimates of specificity and sensitivity were rather precise. The SEs were 0.012 for the specificity and 0.008 for the sensitivity, and corresponding CIs were 0.356–0.402 and 0.955–0.987, respectively.

Table 1. Relationship between sample size and 95% CIs of a proportion (e.g., a sensitivity or specificity).[a]

Sample size	95%CI of a proportion of 0.05	95%CI of a proportion of 0.80
20	0.00–0.25	0.56–0.94
60	0.01–0.14	0.68–0.90
100	0.02–0.11	0.71–0.87
500	0.03–0.07	0.76–0.83
1000	0.04–0.07	0.77–0.82

[a] Selected examples of proportions of 0.05 and 0.8.

POSTERIOR PROBABILITIES (PREDICTIVE VALUES)

A natural question arising after the performance of a diagnostic test is: what is the probability [P(D|Tpos)] that the target disease is present given the index test result? The sensitivity and specificity estimates do not directly answer this question. The probability of target disease given the index test result is a so-called posterior probability, where the prior probability is equal to the prevalence of the disease in the study sample. Quite simply, for a positive test result (Tpos), this probability is estimated by calculating the fraction of TP out of all test result positives:

$$P(\text{D}|\text{Tpos}) = \text{TP}/(\text{TP} + \text{FP}).$$

In the same way, having obtained a negative result (Tneg), we are interested in the probability that the given disease is absent:

$$P(\text{Non-D}|\text{Tneg}) = \text{TN}/(\text{TN} + \text{FN}).$$

The prevalence of disease [P(D)] in the study sample may be regarded as the a priori probability of disease. In the medical field, these posterior probabilities are also known as predictive values *(19)*. Just as with sensitivities and specificities, these posterior disease probabilities depend on the selected cutoff value for a quantitative test. According to Bayes rule, the following relations exist:

$$P(\text{D}|\text{Tpos}) = [\text{Sensitivity} \times \text{P(D)}]/[\text{Sensitivity} \times \text{P(D)} + (1 - \text{Specificity})(1 - \text{P(D)})],$$

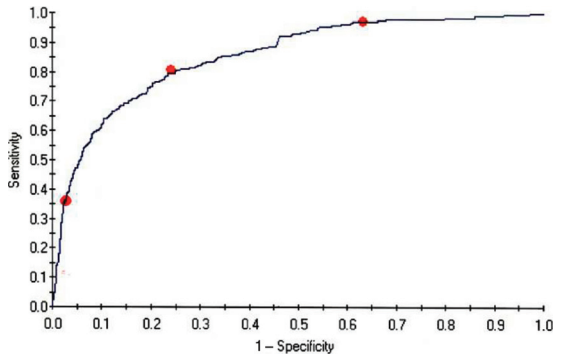

Fig. 4. ROC curve of the D-dimer assay result for diagnosis of DVT in our example study.

The red markers correspond to various cutoff choices (from left to right: 5435 μg/L, 2133 μg/L and 500 μg/L).

$$P(\text{Non-D}|\text{Tneg}) = [\text{Sensitivity} \times (1 - \text{P(D)}]/[\text{Specificity} \times (1 - \text{P(D)}) + \text{P(D)} \times (1 - \text{Specificity})].$$

ROC CURVE

To outline the interdependency of specificity and sensitivity for a given quantitative index test, one may plot the values for all possible cutoff values over the measurement range, which results in the so-called ROC curve, which is shown in Fig. 4 for our D-dimer example *(20–23)*. In the usual plot, sensitivity (*y*) is plotted against (1 − specificity) (*x*) at each possible cutoff value. The better the test, the more the curve is located in the left, upper area. From the curve, a suitable combination of specificity and sensitivity (or accepted FN vs FP proportion) may be selected by a reader, and use the corresponding cutoff for that index test. It could correspond to the maximum of the sum of the specificity and sensitivity or some other point. For the D-dimer example, the traditionally used cutoff of 500 for the D-dimer test corresponds to the point that provides a sensitivity of 0.97 and (1 − specificity) of 0.63.

From this ROC curve a statistical evaluation of an estimated ROC curve area should be performed. The procedure should be related to the way the ROC curve has been estimated, parametrically or nonparametrically, of which the latter approach generally is preferable. Standard computer programs that carry out the task are widely available. Having an SE of the area, one may test whether the area significantly exceeds 0.5, which is the test of whether the index test performs better than chance. Additionally, a 95% CI can be established. Diagnostic tests may also be compared by studying the relationships between their ROC curves. In principle, the test with the largest area under the ROC curve is the best test, although the interpretation becomes more complex if ROC curves of different tests cross each other. A worthless test has an area of 0.5. For the D-dimer example, the area under the ROC curve was 0.86 (SE 0.011), and the 95% CI was 0.84–0.88.

The ROC area provides an overall measure of diagnostic ability. It can be shown that the area under the ROC curve represents, for all possible study pairs of an individual with and without the target disease, the proportion in which individuals with the target disease have a higher (more severe) index test result than individuals without the disease *(20–23)*. ROC curve evaluation may have various advantages, but it also carries some limitations *(23–25)*. Also, the ROC curve does not directly assess the index test performance for a selected cutoff, but can be used for this purpose depending on the desired sensitivity and specificity, or rather the accepted FN and FP proportions; Fig. 4 also shows the sensitivity and specificity of the various D-dimer cutoff points (including 500 μg/L, as well as the cutoff values used in Fig. 5 below).

SELECTION OF CUTOFF VALUE

As can be seen from Fig. 2, the values of specificity and sensitivity of an index test vary inversely with the choice of cutoff point. A suitable tradeoff may be the cutoff point that provides the maximum of the sum of the specificity and sensitivity. This is shown in Fig. 5A for the D-dimer example, where a cutoff close to 2000 μg/L provides a specificity of 0.76 and a sensitivity of 0.80. However, this method of cutoff selection is not necessarily optimal for each situation or acceptable for each user. If an index test is used primarily to rule out the presence of disease, as is the case for the D-dimer assay for exclusion of DVT, the cutoff should be placed at the lower end of the distribution of diseased individuals as shown in Fig. 2, e.g., a cutoff of 500 μg/L. If such a cutoff is selected, the sensitivity becomes almost 1.0, but such a high sensitivity is usually obtained at the cost of a loss of specificity. Depending on the degree of overlap of values, the specificity may become quite low. If, on the other hand, FP results are judged unacceptable, the cutoff should be placed at the upper end of the distribution of values for the nondiseased population. For the D-dimer example, a cutoff corresponding to the 97.5 percentile of the distribution of values for those without DVT (5435 μg/L) provides a specificity of 0.975, but now the sensitivity is only 0.36, i.e., about the reverse of the situation with a cutoff of 500 μg/L (Fig. 5B).

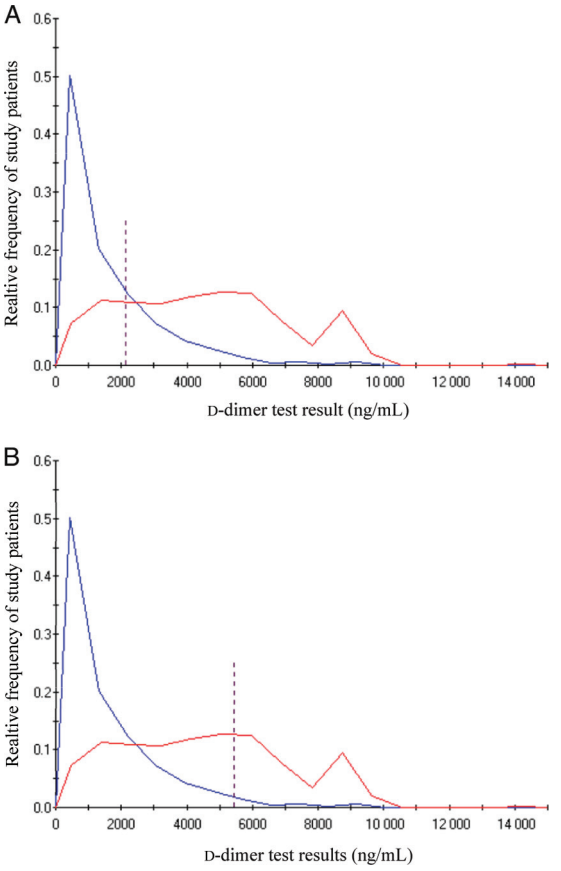

Fig. 5. Alternative choices than 500 μg/L for the cutoff value in the D-dimer example.

(A), Cutoff (2133 μg/L) providing maximum value of the sum of the specificity and sensitivity. (B), Cutoff (5435 μg/L) providing a high specificity value (0.975). Blue line, non-DVT; red line, DVT. The dashed line indicates the cutoff value.

In case the cutoff value is selected in the same study in which sensitivity and specificity of the index test have been estimated, there is a risk of bias. It can be shown that when using the same groups of diseased and nondiseased individuals for estimation of an optimal cutoff point in the data at hand, evaluation of specificity and sensitivity becomes biased *(26, 27)*. A general recommendation is to use independent samples for estimation of the diagnostic cutoff value of the index test and for estimating the corresponding diagnostic accuracy measures. Evaluation of the test in an independent sample also allows the robustness of the test to be assessed.

LIKELIHOOD RATIO

As an alternative to using a cutoff point, one may apply the so-called diagnostic likelihood ratio (LR) principle for interpretation of diagnostic index test results. On the basis of the relative index test result frequency distributions in the nondiseased and diseased groups one may calculate the LR of an index test result (X) as the ratio between the heights of the relative frequency (f) distributions for that specific test result *(28)*. We have:

$$LR(X) = f_D(X)/f_{Non\text{-}D}(X).$$

If the relative frequency of the distribution of diseased individuals exceeds that of the nondiseased individuals, the ratio exceeds 1. This indicates that other factors being equal, disease is more likely than nondisease given the index test result X. More formally, the ratio can be used to calculate posterior probabilities given specific values of X. We have:

$$P(D|X) = P(D) \times LR(X)/[P(D) \times LR(X) + (1 - P(D))],$$

or a more straightforward calculation can be performed using odds instead of probabilities:

$$Odds(D|X) = Odds(D) \times LR(X),$$

using the relation:

$$Odds = P/(1 - P).$$

Odds is an alternative way of expressing probabilities, which is well known from betting games.

The equation states that the posterior odds are equal to the prior odds multiplied by the diagnostic LR for the result X.

For a qualitative test, the following relationships hold true:

$$LR(pos) = Sensitivity/(1 - Specificity),$$

$$LR(neg) = (1 - Specificity)/Sensitivity.$$

Although the concept has been used in various situations, overall the use of diagnostic LRs has been limited in clinical chemistry. Various assumptions are necessary for the concept to be applied in a practical and reliable way. A practical way of deriving the posttest probability of disease from the prevalence (pretest probability of disease) and the diagnostic LR is to apply the Fagan nomogram *(29)*. A recent example is the estimation of the probability of DVT on the basis of testing for D-dimer *(30)*. Finally, it should be mentioned that the diagnostic LR of a result X equals the slope of the ROC curve at the given point.

Comparison of Tests

New diagnostic index tests are commonly compared with an established index test, both of which are com-

pared to a common reference standard. Hence, a statistical assessment is of relevance. When comparing the accuracy of 2 or more diagnostic index tests, a paired design is generally advisable for reasons of both validity and efficiency. In the suspected patient cohort design, the 2 index tests to be compared and the reference standard are carried out on all participants, again independently and blinded for each other's results. With the use of the same nondiseased and diseased individuals (as determined by the reference standard) for both index tests, any bias related to differences in disease spectrum or comorbidity is automatically avoided. Thus, for example, when comparing the sensitivities of 2 index tests, a paired procedure for comparison of proportions such as the McNemar's test should be applied *(31)*. The principle in this statistical procedure is that the number of preferences for index test A (cases detected by index test A but not by index test B) is compared with the number of preferences for index test B, and if the difference exceeds some threshold, 1 index test is significantly better than the other. When comparing 2 diagnostic index tests on the basis of the sensitivities, it is essential that there is no bias with regard to their specificities. If this is the case, the sums of specificity and sensitivity should be compared.

ROC curve areas may also be compared. Here, a paired comparison should also be undertaken when the index tests have been applied on the same groups of individuals. An example of a paired comparison is displayed in Fig. 6. Parametric and nonparametric statistical procedures exist that usually are carried out by computer programs *(22, 32)*, although there may be some limitations, for example when the 2 ROC curves cross *(25)*.

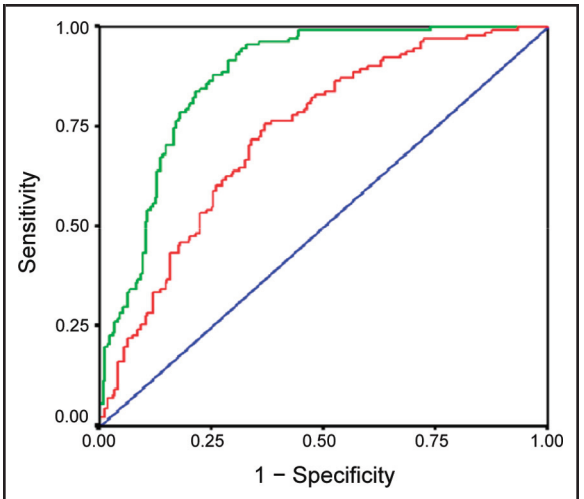

Fig. 6. Comparison of the ROC curves of 2 hypothetical index tests for the same target disease, conducted in the same patients.

The green curve represents a better diagnostic test, both in terms of sensitivity and specificity across all its cutoff points, The blue diagonal represents a worthless test, with equal chance of an FP (1 − specificity) and FN (1 − sensitivity) finding across all cutoff values (i.e. flipping a coin test).

Variations and Limitations of the Accuracy of a Single Test

As mentioned above, the accuracy of a diagnostic test highly depends on the context. Hence, the estimated diagnostic accuracy measures of an index test, regardless of whether they are obtained from data of a suspected patient cohort or from a case control study approach and regardless of what kind of measure (predictive values, sensitivity, specificity, LR, or ROC area) are not constant; they vary across other index test results, patient characteristics, or disease severities *(8, 9, 11)*. We illustrate this for our D-dimer example in Table 2. The overall sensitivity and specificity for the 500 μg/L threshold were 0.97 and 0.37, respectively (upper row). However, when estimating these measures for patient subgroups within the study sample defined by other test results from patient history and physical examination, we found substantial differences in specificity, notably for the malignancy, recent surgery, and pitting-edema subgroups. Using a higher threshold (1000 μg/L) we saw variations in sensitivity as well for, e.g., pregnancy and previous embolism subgroups. The last column of Table 2 shows that this variation in single-test accuracy measures also applies to non–threshold-dependent measures such as the ROC area. The ROC area ranged from 0.79 to 0.98, with 0.86 for the total study group. Although all these differences should not be overinterpreted, the message is that one must always be careful when judging a single test's diagnostic accuracy measures. A diagnostic test should always be placed into a specific clinical context and its results judged on the basis of the diagnostic pathway in which it is to be used *(9–11)*.

Concluding Remarks

A key step in the evaluation of a diagnostic index test is to determine its accuracy, which will indicate the frequency and type of errors that a test will produce when differentiating between patients with and without the disease of interest. The suspected patient cohort design is generally preferable. However, regardless of what design is used or what diagnostic accuracy measure is estimated, there is no such thing as a single accuracy value of a diagnostic test. The predictive values, the sensitivity, specificity, LR, and ROC area for a single test, are not constant but will vary across disease sever-

Table 2. Variations in the sensitivity and specificity (at cut-off values 500 ng/mL and 1000 ng/mL) and the ROC area of the D-dimer test according to various other test results or patient characteristics.

	D-dimer >500		D-dimer >1000		D-dimer (continuous)
	Sensitivity	Specificity	Sensitivity	Specificity	AUC[a] (CI)
Overall	0.97	0.37	0.89	0.55	0.86 (0.84–0.88)
Previous lung embolism					
Yes (n = 173)	1.00	0.37	0.84	0.53	0.82 (0.75–0.90)
No (n = 1913)	0.97	0.37	0.89	0.55	0.86 (0.84–0.88)
Malignancy					
Yes (n = 115)	0.95	0.25	0.95	0.44	0.86 (0.79–0.93)
No (n = 1971)	0.97	0.38	0.89	0.55	0.84 (0.83–0.87)
Recent surgery					
Yes (n = 278)	0.96	0.22	0.90	0.38	0.84 (0.78–0.90)
No (n = 1808)	0.97	0.39	0.89	0.57	0.86 (0.84–0.88)
Leg trauma					
Yes (n = 344)	0.96	0.32	0.85	0.48	0.79 (0.72–0.87)
No (n = 1742)	0.97	0.38	0.89	0.56	0.86 (0.84–0.89)
Pitting edema					
Yes (n = 1301)	0.97	0.32	0.88	0.50	0.84 (0.82–0.87)
No (n = 785)	0.97	0.46	0.90	0.62	0.87 (0.84–0.91)
Pregnancy					
Yes (n = 45)	1.00	0.28	1.00	0.55	0.98 (0.00–1.00)
No (n = 2041)	0.97	0.37	0.89	0.55	0.85 (0.83–0.88)

[a] AUC, area under the ROC curve.

ity, patient characteristics, and other observed test results *(8–11)*. Use of the terminology, characteristics, or properties of a test in itself is thus incorrect, but depends on the context in which the test is used. To improve the reporting of diagnostic accuracy studies, the STARD (STAndards for the Reporting of Diagnostic accuracy) initiative was undertaken several years ago *(2)*. A checklist was developed to guide investigators regarding what information to report on patient recruitment, the order of test execution, and the number of patients undergoing the test under evaluation, the reference test, or both *(2)*. Diagnosis in practice is often about combining results from multiple tests or about the added value of a new test beyond what is already known. These issues will be addressed in our second report. Owing to the various pitfalls in diagnostic accuracy evaluations, several evaluations in different places may be necessary to provide a reliable indication of the performance of a given test. Systematic reviews of diagnostic accuracy studies may provide additional insights, as addressed in our third report. Finally, proper identification of the target disease by an accurate or less invasive index test does not automatically translate into improved decision-making, let alone patient benefits. Diagnostic accuracy should therefore be seen only as an intermediate outcome, albeit a good one, but still an intermediate outcome. In report 4 of our series we will discuss study designs that aim to directly measure downstream consequences due to testing on outcomes relevant for patients or healthcare.

Author Contributions: *All authors confirmed they have contributed to the intellectual content of this paper and have met the following 3 requirements: (a) significant contributions to the conception and design, acquisition of data, or analysis and interpretation of data; (b) drafting or revising the article for intellectual content; and (c) final approval of the published article.*

Authors' Disclosures or Potential Conflicts of Interest: *Upon manuscript submission, all authors completed the author disclosure form. Disclosures and/or potential conflicts of interest:*

Employment or Leadership: None declared.
Consultant or Advisory Role: None declared.
Stock Ownership: None declared.
Honoraria: None declared.
Research Funding: The Netherlands Organisation for Health Research and Development ZonMWK, European Commission Grant

FP7; K.G.M. Moons, the Netherlands Organisation for Scientific Research (project 9120.8004 and 918.10.615).
Expert Testimony: None declared.

Role of Sponsor: The funding organizations played no role in the design of study, choice of enrolled patients, review and interpretation of data, or preparation or approval of manuscript.

References

1. Linnet K, Boyd JC. Selection and analytical validation of methods – with statistical techniques. In: Burtis C, Ashwood ER, Bruns D, eds. Tietz textbook of clinical chemistry and molecular diagnostics. 5th ed. New York: Elsevier; 2012. p 7–47.
2. 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 intiative. Clin Chem 2003; 49:1–6.
3. Linnet K. A review on the methodology for assessing diagnostic tests. Clin Chem 1988;34: 1379–86.
4. Oudega R, Moons KG, Hoes AW. Ruling out deep venous thrombosis in primary care. A simple diagnostic algorithm including D-dimer testing. Thromb Haemost 2005;94:200–5.
5. Toll DB, Oudega R, Bulten RJ, Hoes AW, Moons KG. Excluding deep vein thrombosis safely in primary care. J Fam Pract 2006;55:613–8.
6. Moons KG, Grobbee DE. When should we remain blind and when should our eyes remain open in diagnostic research. J Clin Epidemiol 2002;55: 633–6.
7. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. N Engl J Med 1978;299:926–30.
8. Hlatky MA, Pryor DB, Harrell FE Jr, Califf RM, Mark DB, Rosati RA. Factors affecting sensitivity and specificity of exercise electrocardiography. Multivariable analysis. Am J Med 1984;77: 64–71.
9. Moons K, van Es GA, Deckers JW, Habbema JD, Grobbee DE. Limitations of sensitivity, specificity, likelihood ratio, and Bayes' theorem in assessing diagnostic probabilities: a clinical example. Epidemiology 1997;8:12–7.
10. Leeflang MM, Bossuyt PM, Irwig L. Diagnostic test accuracy may vary with prevalence: implications for evidence-based diagnosis. J Clin Epidemiol 2009;62:5–12.
11. Moons KG, van Es GA, Michel BC, Buller HR, Habbema JD, Grobbee DE. Redundancy of single diagnostic test evaluation. Epidemiology 1999; 10:276–81.
12. Reitsma JB, Rutjes AW, Khan KS, Coomarasamy A, Bossuyt PM. A review of solutions for diagnostic accuracy studies with an imperfect or missing reference standard. J Clin Epidemiol 2009;62: 797–806.
13. de Groot JA, Janssen KJ, Zwinderman AH, Bossuyt PM, Reitsma JB, Moons KG. Correcting for partial verification bias: a comparison of methods. Ann Epidemiol 2011;21:139–48.
14. Rutjes AW, Reitsma JB, Vandenbroucke JP, Glas AS, Bossuyt PM. Case-control and two-gates design in diagnostic accuracy studies. Clin Chem 2005;51:1335–41.
15. Biesheuvel CJ, Vergouwe Y, Oudega R, Hoes AW, Grobbee DE, Moons KG. Advantages of the nested case-control design in diagnostic research. BMC Med Res Methodol 2008;8:48.
16. Lijmer JG, Moi BW, Heisterkamp S, Bonsel GJ, Prins MH, Van der Meulen JHP, Bossuyt PMM. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA 1999;282:1061–6.
17. Brown LD, Cai TT, DasGupta A. Interval estimation for a binomial proportion. Stat Science 2001; 16:101–33.
18. Bachmann LM, Puhan MA, ter Riet G, Bossuyt PM. Sample sizes of studies on diagnostic accuracy: literature survey. BMJ 2006;332: 1127–9.
19. Vecchio TJ. Predictive value of a single diagnostic test in unselected populations. N Engl J Med 1966;274:1171–3.
20. Metz CE, Goodenough DJ, Rossmann K. Evaluation of receiver operating characteristic curve data in terms of information theory, with applications in radiography. Radiology 1973;109:297–303.
21. Hanley JA, Mcneil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143:29–36.
22. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 1983;148:839–43.
23. Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clin Chem 1993;39: 561–77.
24. Obuchowski NA, Lieber ML, Wians FH Jr. ROC curves in clinical chemistry: uses, misuses, and possible solutions. Clin Chem 2004;50:1118–25.
25. Moons KG, Stijnen T, Michel BC, Büller HR, Van Es GA, Grobbee DE, Habbema JD. Application of treatment thresholds to diagnostic-test evaluation: an alternative to the comparison of areas under receiver operating characteristic curves. Med Decis Making 1997;17:447–54.
26. Linnet K, Brandt E. Assessing diagnostic tests once an optimal cutoff point has been selected. Clin Chem 1986;32:1341–6.
27. Leeflang MM, Moons KG, Reitsma JB, Zwinderman AH. Bias in sensitivity and specificity caused by data-driven selection of optimal cut-off values: mechanism, magnitude, and solutions. Clin Chem 2008;54:729–37.
28. Albert A. On the use and computation of likelihood ratios in clinical chemistry. Clin Chem 1982; 28:1113–9.
29. Fagan TJ. Letter: Nomogram for Bayes theorem. N Engl J Med 1975;293:257.
30. Geersing GJ, Toll DB, Janssen KJM, Oudega R, Blikman MJC, Wijland R, et al. Diagnostic accuracy and user-friendliness of 5 point-of-care D-dimer tests for the exclusion of deep vein thrombosis. Clin Chem 2010;56:1758–66.
31. Altman DG. Practical statistics for medical research. 1st ed. London: Chapman & Hall; 1991. p 258.
32. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating curves: a nonparametric approach. Biometrics 1988;44:837–45.