Chapter 18

Evaluation of Diagnostic Tests

Brendan J. Barrett and John M. Fardy

Abstract

As technology advances, diagnostic tests continue to improve and each year, we are presented with new alternatives to standard procedures. Given the plethora of diagnostic alternatives, diagnostic tests must be evaluated to determine their place in the diagnostic armamentarium. The first step involves determining the accuracy of the test, including the sensitivity and specificity, positive and negative predictive values, likelihood ratios for positive and negative tests, and receiver operating characteristic (ROC) curves. The role of the test in a diagnostic pathway has then to be determined, following which the effect on patient outcome should be examined.

Key words Diagnostic tests, Sensitivity, Specificity, Positive predictive value, Negative predictive value, Likelihood ratio, Receiver operating characteristic curve

1 Introduction

Diagnostic tests are used to increase the likelihood of the presence or absence of illness, to provide prognostic information and, in some situations, to predict a response to treatment. The ability of a diagnostic test to identify a potential underlying disorder depends not only on the characteristics of the test itself, but also on the particular situation in which it is used. The prevalence of the disease in the population, the specifics of the population studied, and the spectrum of the disease being sought may influence the way a diagnostic test performs. In this chapter, the characteristics of diagnostic tests are examined together with how these characteristics can be used to choose the most useful diagnostic tests. It should be noted at the outset that the term "diagnostic test" is not limited to laboratory or imaging studies, but can also include parts or all of a clinician's assessment. For example, Ehrenstein et al. examined the ability of rheumatologists to diagnose rheumatoid arthritis based on clinical assessment alone [1], while a meta-analysis was used to assess the accuracy of the Lever sign to diagnose anterior cruciate ligament tear [2]. The same methods to assess

Patrick S. Parfrey and Brendan J. Barrett (eds.), *Clinical Epidemiology: Practice and Methods*, Methods in Molecular Biology, vol. 2249, https://doi.org/10.1007/978-1-0716-1138-8_18,

accuracy can even be applied to patient-reported measures as was done for the prediction of mucosal inflammation in inflammatory bowel disease [3].

In order to determine the accuracy of a diagnostic test, an arbiter is necessary to decide whether the test result is correct or not. This is known as the "gold standard" or reference standard. In some instances, the "gold standard" is an established test or combination of tests which confirms the diagnosis, while in other cases the "gold standard" requires follow-up over time to confirm or refute the diagnosis. When considering the characteristics of a diagnostic test, one must consider the "gold standard" to which it is compared and determine whether or not it is an appropriate one. The comparison to the "gold standard" should be carried out in a blinded fashion so as to prevent bias in the interpretation of the diagnostic test or the reference standard. Further issues in the design of diagnostic accuracy studies are discussed in a later section.

In the past, assessment of diagnostic tests might have been limited to studies of accuracy, but it is now well recognized that tests form part of a diagnostic pathway. Test results are used to alter the probability of diagnoses in the context of what is already known about the case and the results of other tests that might have been completed at the same time. There is recognition that the results of groups of tests may not be independent. As such, the specific contribution of a particular test needs to be determined. This has been discussed by Moons and colleagues, where the information gain from adding a test can be quantified in terms of an increase in the area under the ROC curve (see below), net reclassification improvement, or by decision curve analysis [4]. Furthermore, following studies of the clinical validity of a test, the clinical utility of the test then needs to be established [5]. There has been plenty of literature on the best approach to assessing the clinical utility of tests. Such evaluations may include learning about the full range of effects of tests on patients: psychological, behavioral, and social effects together with the impact of subsequent therapies on longer term health outcomes and costs [6, 7].

2 Diagnostic Test Accuracy Criteria

2.1 Sensitivity and Specificity

The classic parameters used to characterize a diagnostic test are the sensitivity and specificity of the test. The sensitivity of a test refers to its ability to identify persons with the disease. It can be defined as "the proportion of people who truly have a designated disorder who are so identified by the test" [8]. A very sensitive test is one that identifies most people with the disorder in question. A test which is very sensitive is prone to false-positive results, that is, it may incorrectly label people as having the disease when, in fact, they do not have it.

Table 1
Assessment of diagnostic tests using 2 × 2 contingency table

New test	Gold standard			
		Positive	**Negative**	**Total**
	Positive	True positive a	b False positive	a + b
	Negative	False negative c	d True negative	c + d
	Total	a + c	b + d	

Sensitivity: a/(a + c)
Specificity: d/(b + d)
Positive predictive value: a/(a + b)
Negative predictive value: d/(c + d)

The specificity of a test, on the other hand, refers to its ability to correctly identify the disease in question. It can be defined as "the proportion of people who are truly free of a designated disorder who are so identified by the test" [8]. A very specific test would be unlikely to incorrectly label an individual as having the disorder in question if, in fact, they do not have the disorder. However, a test that is very specific is more prone to false-negative results, that is, it may fail to identify the disease in some persons who actually have it.

There is always a trade-off between sensitivity and specificity; as one increases, the other tends to decrease [9]. The higher the cut-off used to say a test is positive, the more specific the test becomes, but this higher specificity comes at a price. As the cut-off is increased, the sensitivity decreases and the test is more likely to miss affected individuals. In some situations, such as screening for a disease, a lower cut-off might be used to create a very sensitive test so as not to miss anyone with the disorder in question. In other situations, when using a test to confirm a diagnosis, a higher cut-off making the test highly specific would be more desirable so as not to incorrectly label anyone with the disorder.

The sensitivity and specificity of a diagnostic test can be calculated using information obtained by comparing the performance of a diagnostic test to a gold standard or reference standard. Typically, these results are summarized in a 2 × 2 contingency table as shown in Table 1. Such tables can of course be extended to illustrate the distribution of data at different test cut-offs. Sensitivity and specificity are not directly influenced by disease prevalence, but are affected by the disease severity spectrum. A test that is sensitive for detection of advanced disease may be less sensitive for detection of earlier stages. An example would be the chest X-ray for detection of lung cancer.

2.2 Positive and Negative Predictive Values

The sensitivity and specificity of a diagnostic test are useful to describe how well a test performs, but they do not give us much information on the significance of a positive or negative test for an

individual patient. This information can be obtained from the positive and negative predictive values of the test. The positive predictive value describes "the proportion of people with a positive test who have the disease" [9]. Similarly, the negative predictive value describes "the proportion of people with a negative test who are free of disease" [9]. These ratios are calculated across the table rather than down the table using the formulae in Table 1. These parameters are more useful to the clinician and the patient as they give the predictive value of a positive and a negative test. A test with a high positive predictive value makes the disease quite likely in a subject with a positive test. A test with a high negative predictive value makes the disease quite unlikely in a subject with a negative test.

Although the positive and negative predictive values of a test are intuitively more useful to the clinician and patient, the predictive values are less stable and are dependent on the prevalence of disease. This makes them less portable from population to population. It also means that positive and negative predictive values derived from a study may not apply to any given patient if that patient's pre-test probability of disease differs from the prevalence of the disease in the study sample.

2.3 Case Study

Let us take a hypothetical new test used to rapidly detect an infectious process usually diagnosed by a culture technique, which may take up to a month to provide a result (this is the case for several newer tests for tuberculosis). In a cohort of affected and unaffected subjects in which the prevalence of disease is 50%, how does the new test compare to the culture technique? The results in Table 2 show a new test with excellent sensitivity and good specificity. This test would be a good screening test and a reasonable confirmatory test. The positive predictive value of 82% and the negative predictive value of 88% suggest the new test is quite beneficial to patients and doctors.

However, if the prevalence of the disease is 10% instead of 50%, and the sensitivity and specificity are the same, the positive and

Table 2
Assessment of a new diagnostic test when prevalence of disease is 50%

New test	Gold standard			
		Positive	Negative	Total
	Positive	45 (a)	10 (b)	55
	Negative	5 (c)	40 (d)	45
	Total	50	50	100

Prevalence of disease = 50/100 = 50%
Sensitivity = a/(a + c) = 45/50 = 90%
Specificity = d/(b + d) = 40/50 = 80%
Positive predictive value = a/(a + b) = 45/55 = 82%
Negative predictive value = d /(c + d) = 40/45 = 88%

Table 3
Assessment of a new diagnostic test when prevalence of disease is 10%

New test	Gold standard			
		Positive	Negative	Total
	Positive	45 (a)	90 (b)	135
	Negative	5 (c)	360 (d)	365
	Total	50	450	500

Prevalence of disease = 50/500 = 10%
Sensitivity = a/(a + c) = 45/50 = 90%
Specificity = d/(b + d) = 360/450 = 80%
Positive predictive value = a/(a + b) = 45/135 = 33%
Negative predictive value = c/(c + d) = 360/365 = 98%

negative predictive values change as shown in Table 3. Although the negative predictive value has increased from 88% to 98%, the positive predictive value has dropped to 33%. This test, which was initially a very good predictor of disease when prevalence was 50%, has much poorer positive predictive value when the disease prevalence drops to 10%. In fact, with lower disease prevalence, the test produces twice as many false positives as true positives. In general, diagnostic tests will function most efficiently when the prevalence (or pre-test probability) is between 40% and 60% and provide much less information at the extremes of pre-test probability [9]. This points to a general point about tests: when you already are pretty sure that a patient either does or does not have the diagnosis in question, additional testing may not alter that probability very much.

2.4 Likelihood Ratios

The ideal test parameter would be one that has predictive value and is stable with changes in prevalence. The likelihood ratio is such a parameter. A likelihood ratio expresses the relative odds that a given level of a diagnostic test result would be expected in a patient with (as opposed to one without) the target disorder [8]. As with the other parameters, likelihood ratios are calculated from the 2 × 2 table.

Likelihood ratio for a positive test.

$$LR+ = (a/a + c)/(b/b + d) = \text{sensitivity}/(1\text{-specificity})$$

Likelihood ratio for a negative test.

$$LR- = (c/a + c)/(d/b + d) = (1\text{-sensitivity})/\text{specificity}$$

Because the likelihood ratios are calculated from the sensitivity and specificity, they are also stable with changes in prevalence of disease. The predictive value of the likelihood ratio calculates the post-test odds of disease from the pre-test odds of disease using the following formula:

$$\text{Post-test odds} = \text{Pre-test odds} \times LR+$$

The pre-test odds of disease is similar to the pre-test probability of disease and can be calculated with the following formula:

Pre-test odds = Pretest probability/(1 – Pre-test probability)

The pre-test probability of disease is usually estimated from the clinical information or from published reports.

A diagnostic test with likelihood ratios near unity does not have much effect on the post-test probability of disease and therefore is not very useful for decision-making. On the other hand, very large LR+ or very small likelihood LR- ratios have a significant impact on the post-test probability of disease. An LR for a positive test of 10 or more means that a positive test is good at ruling in a diagnosis while an LR for a negative test of 0.1 or less means that a negative test is good at ruling out a diagnosis [10]. Likelihood ratios between 5–10 if test positive or 0.1–0.2 if test negative lead to moderate changes in the post-test probability while those between 2–5 (0.2–0.5) lead to smaller changes.

The use of likelihood ratios to characterize diagnostic tests highlights the importance of the pre-test probability of disease in the performance of a diagnostic test. If the pre-test probability of disease is very high or very low, a diagnostic test will have to be very good to make a significant difference in the post-test probability of disease. Diagnostic tests will perform best when the pre-test probability of disease is about 50% and generally will perform less well at the extremes of pre-test probability [9]. If the pre-test probability of disease is so high or so low as to rule in or rule out a diagnosis, a diagnostic test is not warranted [10]. This statistical approach to modifying prior probabilities (or odds) in light of new information is Bayesian.

2.5 Overall Test Accuracy

These various parameters used to characterize diagnostic tests can help in choosing one test over another, but they do not provide a summary estimate of the accuracy of the test. The receiver operating characteristic (ROC) curve can be used for this purpose. An ROC curve is a plot of test sensitivity (plotted on the y-axis) versus its false-positive rate (1 – specificity) (plotted on the x-axis) [11]. As the cut-off value for a positive test is moved up or down, the sensitivity and specificity of the test change. Figure 1 is an example of an ROC curve for a hypothetical diagnostic test. In this example, raising the cut-off value would lead to high specificity and low sensitivity, with coordinates toward the lower left-hand corner of the curve. Lowering the cut-off value for a positive test would lead to a progressive increase in sensitivity and a progressive decrease in specificity moving up along the curve toward the upper right-hand corner. The point on the curve closest to the upper left-hand corner (which represents 100% sensitivity and 100% specificity) would represent the cut-off value which offers the best balance between sensitivity and specificity. This may not always be the best

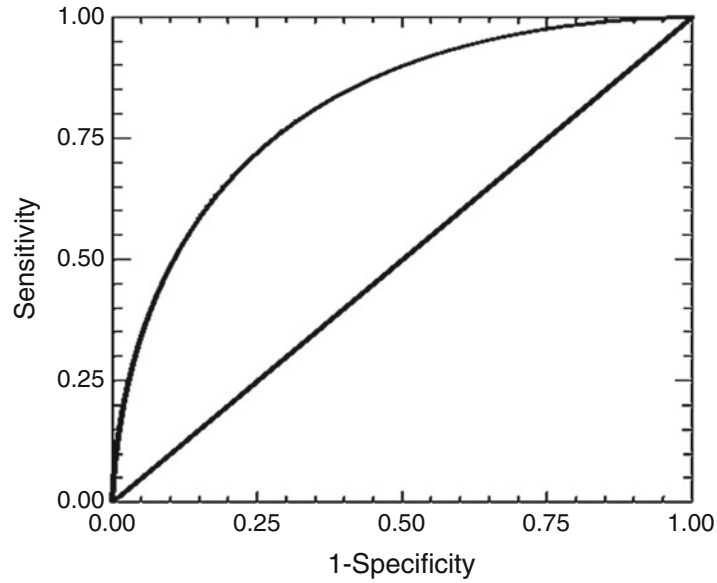

Fig. 1 Receiver operating characteristic (ROC) curve for assessing diagnostic tests

cut-off to choose, depending on the purpose of the test. For a screening test, sensitivity would be favored over specificity, while for a confirmatory test specificity would be favored over sensitivity. In general, one needs to consider the clinical impact of false-positive and false-negative test results and weigh these against each other to determine the most useful cutoff for any given context.

The ROC curve also provides information on the overall accuracy of the diagnostic test. The area under the ROC curve (the area to the right of the curved line in Fig. 1) is a popular measure of the accuracy of a diagnostic test [11]. The ROC curve area can take on values between 0.0 and 1.0, with an area of 1.0 representing a perfectly accurate test. A test with an area of 0.0 is perfectly inaccurate; all patients with the disease have negative results, while all those without the disease have positive results. Such a test would have perfect accuracy if the interpretation of the test were reversed. Therefore, the practical lower bound for the area under the ROC curve is 0.5, which is bounded by the straight line from coordinates 0–11. This line is known as the chance diagonal on an ROC plot [11]. The area under the ROC curve can be used to compare the accuracy of diagnostic tests. It should be noted that in a given study the area under the curve is an estimate with an associated standard error. This can be used to calculate confidence intervals around the estimated area and is also used when the areas under the ROC curves associated with different tests are being compared. Both parametric and non-parametric statistical procedures exist to

compare areas under ROC curves, including adjustments for paired samples if the two tests being compared were completed within the same subjects [12, 13].

If the concern is the accuracy of a test, the percentage of patients correctly classified by the test under evaluation can be assessed. In Table 1, accuracy can be calculated as follows:

Accuracy: (a+d)/(a+b+c+d)

Unfortunately, the overall accuracy is highly dependent on the prevalence of the disease. Another option for a single indicator of test performance is the diagnostic odds ratio (DOR). This is the ratio of the odds of positivity in the diseased relative to the odds of positivity in the non-diseased [14]. Like the odds ratio in any 2×2 table it is calculated using the following formula:

$$DOR = ad/bc$$

There is also a close relationship between the DOR and the likelihood ratios:

$$DOR = LR+ / LR- \text{ [14]}$$

The value of the DOR ranges from 0 to infinity with higher values associated with better performance of a diagnostic test. A value of 1 suggests that a test does not discriminate well between those with and without the target disorder, while values lower than 1 suggest improper interpretation of the diagnostic test (more negative tests among the diseased). As with likelihood ratios, the DOR is not dependent on the prevalence of disease, but like sensitivity and specificity is influenced by the disease spectrum in the study population [14]. The DOR can also be useful in meta-analysis of diagnostic studies.

In all of the previous discussion, it has been assumed that the reference or "gold" standard will yield a binary outcome of disease presence or absence. However, this is not always the case, as for example when echocardiographically determined left ventricular mass as a continuous measure serves as the reference standard when evaluating features of the ECG as a diagnostic test. In that case, a different statistical approach has been proposed for estimating sensitivity, specificity, and the ROC curve [15]. An alternate approach using information theoretical concepts also permits consideration of quantitative reference results while explicitly taking into account variation in pre-test probabilities [16].

In addition, the reference standard itself may not always be perfect, and in that situation the use of Bayesian latent class models can allow evaluation of novel tests [17–19]. A web-based application has been developed to allow the less statistically accomplished researcher to complete the required analyses via a user-friendly interface [20].

3 Design of Diagnostic Accuracy Studies

Given the various tools available, how would one set out to evaluate a new diagnostic test? The criteria have been discussed in standard textbooks of clinical epidemiology and are outlined below [8, 9]. These criteria center around a blinded evaluation of the new test versus a "gold standard" in an appropriate population. The reproducibility and the interpretation of the test should be standardized and the test procedure should be well described. Finally, the clinical utility should be documented.

The importance of a blinded evaluation of the diagnostic test versus the reference standard is paramount in the evaluation of a new diagnostic test. Knowledge of the results of either the diagnostic test or the reference standard could lead to bias when interpreting the results of the other. Lack of a blinded comparison would invalidate the results of the study.

The population chosen for study is also a critical factor in the assessment of a diagnostic test. Test performance will vary with disease prevalence and with disease severity, such that diagnostic test performance often varies across population subgroups [21, 22]. The sample population chosen for evaluation of the diagnostic test should be similar to the population for which the test is intended, in terms of both the prevalence and severity of the disease. The comparison group should be comprised of individuals from that group, those suspected of having the target disorder but not actually having the disease as opposed to "normal" individuals. In essence, the test should be evaluated under the same conditions in which it will be used. Assessing test accuracy in samples selected to include cases with obvious or severe disease as well as healthy controls will tend to overestimate the accuracy of the test under routine conditions. Similarly, it would be inappropriate to exclude cases from a study of test accuracy post hoc based on the results of the gold standard as doing so will lead to biased estimates of test accuracy under field conditions [23].

In studies of the accuracy of diagnostic tests, it is important that all members of the sample population undergo both the test being assessed as well as the "gold standard." In a systematic review of the sources of bias and variation in diagnostic test accuracy studies, Whiting and colleagues found that use of a case–control design, observer variability, availability of clinical information, choice of reference standard, disease prevalence and severity as well as verification biases were the major sources with generally greater impact on the estimate of sensitivity than specificity [24]. Methods for determining sample size for studies of the accuracy of diagnostic tests are tailored to the particular indices that are being studied. Sample size estimates can be calculated for several accuracy indices including sensitivity and specificity, the area under the receiver

operating characteristic curve, the sensitivity at a fixed false-positive rate, and the likelihood ratio [25].

The reproducibility of the test should also be evaluated particularly when it involves a subjective interpretation of the results. Both the inter-observer and intra-observer variation should be examined and evaluated with an appropriate measure, such as a kappa statistic, which reveals the degree of agreement between test readers. The test procedure should be well described so that it can be replicated by others. As well, there may be a significant learning curve associated with the interpretation of a new diagnostic test and this must be taken into account as the test is evaluated.

Given the plethora of studies that may exist evaluating the accuracy of a given diagnostic test, there has been interest in completion of diagnostic test accuracy systematic reviews and meta-analyses. The challenges involved have been addressed by a Cochrane Methods group [26]. A systematic review of the major sources of bias has been reported [18]. Tools were developed to assess the quality of the constituent studies [27, 28]. Challenges are often posed by heterogeneity in the design, setting, and results of the various primary studies. Care needs to be taken when formulating the questions for the systematic review. A PRISMA guideline recommending preferred reporting items in systematic reviews of diagnostic test accuracy studies has been published [29], but evaluation of recently published systematic reviews shows that reports are still not fully informative [30].

Finally, there have been many reports in the past few years where machine learning or other deep learning techniques have been applied to development or assessment of diagnostic tests [31–35]. Technology is also creeping into diagnostic processes with the use of Smartphone and Computer-assisted techniques [36, 37].

4 Factors Relevant to the Choice of Diagnostic Tests

The choice of diagnostic tests is certainly influenced by test performance, but this is not the only important factor to be considered. Although a Ferrari may outperform the competition, its cost and seating capacity may make it unsuitable for the job at hand. In choosing a diagnostic test, one must consider, in addition to test performance, the cost, availability, acceptability, and utility of the diagnostic test. A practical hierarchy can be defined based on (1) diagnostic power or performance, (20 availability and acceptability where considered relevant, and (3) cost [38]. When several test options exist for a given scenario, network meta-analysis has been used to compare them at different test cut-offs [39]. Evaluation of costs alongside clinical impacts can be done using cost-effectiveness approaches [40].

Cost and availability are obvious concerns when one considers the choice of diagnostic tests. A very expensive test with limited availability would have to outperform standard tests by a wide margin before it could be considered for routine use. The acceptability of the diagnostic test is also a major concern, particularly for the patient. An invasive test with potentially serious complications will not be accepted readily by patients, particularly if there is a safer, non-invasive alternative. One must also consider that information produced in research about diagnostic tests is utilized by several different types of decision makers who are interested in different types of information [41]. Policy-making organizations will be more concerned with the "evidence-based" assessment and cost of testing, while patients may place more emphasis on anecdotal experience and the reassurance value of testing. Physicians will typically find themselves acting as representatives of the medical profession and its body of knowledge, and as advocates for each patient [41].

The final arbiter in the choice of diagnostic tests is the clinical utility of the test under scrutiny. Studies of diagnostic test accuracy may, on their own, provide sufficient information to infer clinical value if a new diagnostic test is safer or more specific than the old test, provided both are of similar sensitivity and that treatment based on results of the old test has been shown to improve patient outcomes in clinical trials [42]. Establishing whether a new test improves patient outcomes beyond the outcomes achieved using an older test or maybe no test prior to treatment may require the completion of randomized trials [43]. A randomized trial can assess the outcomes of patients undergoing testing, document adverse effects, and assess impact on management decision-making and measure patient satisfaction and the cost-effectiveness of testing [44]. A variety of randomized designs have been proposed with the choice among them depending upon the objective of testing and whether alternative test/treat strategies are to be compared [45]. A framework for evaluating the links and mechanisms whereby outcomes are impacted in diagnostic test/treat trials has been proposed [46]. Concerns have been raised about the efficiency of some designs proposed for test/treat trials. It has been suggested that in the case where two tests are being compared in terms of clinical utility, a paired design in which each participant undergoes both tests with subsequent treatment only randomly assigned when the test results are discordant may be more efficient [47]. Sample size formulae for binary and continuous outcomes have also been proposed by the same authors [47]. Ethical issues that arise in relation to these trials include the need for equipoise, not so much with regard to the relative accuracy of tests, but rather with regard to the comparative health impact of alternative test/treat strategies. In addition, if a clustered design is followed, there is a need for those who decline participation to be aware that the

whole diagnostic process in a particular clinic or hospital, for example, may be influenced by the assignment of that site to a novel test/treat strategy for the trial [48].

As commonly done in economic analyses, decision models can also be used to compare various test/treat strategies, but the results depend critically on the accuracy of the assumptions and estimates used to build and inform the models [6].

5 Combinations of Diagnostic Tests

Diagnostic tests are often applied in groups or profiles or may be used sequentially to improve precision in diagnosis. Studies of test combinations have to consider the non-independence of the results. Novielli and colleagues have described a framework to do that in relation to diagnosis of deep vein thrombosis [49]. Similarly, various combinations of liquid-based cytology and human papilloma virus testing have been evaluated for screening for cervical in-situ neoplastic lesions [50].

6 Conclusion

Diagnostic test performance can be assessed using a number of different measures which assess the accuracy and predictive value of the tests. The choice of diagnostic tests, however, is more complex than a simple assessment of performance, and consideration of broader issues such as patient outcomes, acceptability and cost-effectiveness of testing is necessary. By using the appropriate criteria to assess diagnostic test performance, followed by randomized trials to measure clinical utility, the choice of the best diagnostic test to solve a diagnostic problem can be made.

References

1. Ehrenstein B, Pongratz G, Fleck M, Hartung W (2018) The ability of rheumatologists blinded to prior workup to diagnose rheumatoid arthritis only by clinical assessment: a cross-sectional study. Rheumatology 57:1592–1601
2. Reiman MP, Reiman CK, Decary S (2018) Accuracy of the lever sign to diagnose anterior cruciate ligament tear: a systematic review with meta-analysis. Int J Sports Phys Ther 13 (5):774–788
3. De Jong MJ, Roosen D, Degens JHRJ, van den Heuvel TRA, Romberg M, Hameeteman W, Bodelier AGL, Romanko I, Lukas M, Winkens B, Markus T, Masclee AAM, van Tubergen A, Jonkers DMAE, Pierik MJ (2018) Development and validation of a patient-reported score to screen for mucosal inflammation in inflammatory bowel disease. J Crohns Colitis 13(5):555–563. https://doi.org/10.1093/ecco-jcc/jjy196
4. Moons KGM, deGroot JAH, Linnet K, Reitsma JB, Bossuyt PMM (2012) Quantifying the added value of a diagnostic test or marker. Clin Chem 58(10):1408–1417
5. Linnet K, Bossuyt PMM, Moons KGM, Reitsma JB (2012) Quantifying the accuracy of a diagnostic test or marker. Clin Chem 58 (9):1292–1301

6. Bossuyt PMM, Reitsma JB, Linnet K, Moons KGM (2012) Beyond diagnostic accuracy: the clinical utility of diagnostic tests. Clin Chem 58 (12):1636–1643
7. Atkin W, Cross AJ, Kralj-Hans I, MacRae E, Piggott C, Pearson S, Wooldrage K, Brown J, Lucas F, Prendergast A, Marchevsky N, Patel B, Pack K, Howe R, Skrobanski H, Kerrison R, Swart N, Snowball J, Duffy SW, Morris S, von Wagner C (2019) Halloran S Faecal immunochemical tests versus colonoscopy for post-polypectomy surveillance: an accuracy, acceptability and economic study. Health Technol Assess 23(1):1–84
8. Guyatt G, Drummond R, Meade MO, Cook DJ (eds) (2008) Users' guides to the medical literature: a manual for evidence-based clinical practice, 2nd edn. New York, McGraw Hill
9. Haynes RB, Sackett DL, Guyatt GH, Tugwell P (2005) Clinical epidemiology: how to do clinical practice research, 3rd edn. Lippincott, Williams and Wilkins, Philadelphia, PA
10. Grimes D, Schulz K (2005) Refining clinical diagnosis with likelihood ratios. Lancet 365:1500–1505
11. Obuchowski NA (2003) Receiver operating characteristic curves and their use in radiology. Radiology 229:3–8
12. Hanley JA, McNeil BJ (1983) A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 148:839–843
13. DeLong ER, DeLong DM, Clarke-Pearson DL (1988) Comparing the areas under two or more correlated receiver operating curves: a nonparametric approach. Biometrics 44:837–845
14. Glas SG, Lijmer JG, Prins MH et al (2003) The diagnostic odds ratio: a single indicator of test performance. J Clin Epidemiol 56:1129–1135
15. Shiu S-Y, Gatsonis C (2012) On ROC analysis with nonbinary reference standard. Biom J 54 (4):457480
16. Reibnegger G (2013) Beyond the 2x2 contingency table: a primer on entropies and mutual information in various scenarios involving m diagnostic categories and n categories of diagnostic tests. Clin Chim Acta 425:97–103
17. Joseph L, Gyorkos TW, Coupal L (1995) Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. Am J Epidemiol 141(3):263–272
18. Limmathurotsakul D, Turner EL, Wuthiekanun V, Thaipadungpanit J, Suputtamongkol Y, Chierakul W et al (2012) Fool's gold: Why imperfect reference tests are undermining the evaluation of novel diagnostics: A reevaluation of 5 diagnostic tests for leptospirosis. CID 55:322–331
19. Pan-ngum W, Blacksell SD, Lubell Y, Pukrittayakamee S, Bailey MS, deSilva HJ et al (2013) Estimating the true accuracy of diagnostic tests for Dengue infection using Bayesian latent class models. PLoS One 8(1):1–7
20. Lim C, Wannapinij P, White L, Day NPJ, Cooper BS, Peacock SJ et al (2013) Using a web-based application to define the accuracy of diagnostic tests when the gold standard is imperfect. PLoS One 8(11):1–8
21. Mullherin SA, Miller MC (2002) Spectrum bias or spectrum effect? Subgroup variation in diagnostic test evaluation. Ann Int Med 137:598–602
22. Goudsmit M, van Campen J, Schilt T, Hinnen C, Franzen S, Schmand B (2018) One size does not fit all: Comparative diagnostic accuracy of the Rowland Universal Dementia Assessment Scale and the Mini mental State Examination in a memory clinic population with very low education. Dement Geriatr Cogn Disord Extra 8:290–305
23. Lang S, Armstrong N, Deshpande S, Ramaekers B, Grimm S, de Kock S, Kleijnen J, Westwood M (2019) Clinically inappropriate post hoc exclusion of study participants from test accuracy calculations: the ROMA score, an example from a recent NICE diagnostic assessment. Ann Clin Biochem 56(1):72–81
24. Whiting PF, Rutjes AWS, Westwood ME, Mallett S, and the QUADAS-2 Steering Group (2013) A systematic review classifies sources of bias and variation in diagnostic test accuracy studies. J Clin Epidemiol 66:1093–1104
25. Obuchowski NA (1998) Sample size calculations in studies of test accuracy. Stat Meth Med Res 7:371–392
26. Leeflang MMG, Deeks JJ, Takwoingi Y, Macaskill P (2013) Cochrane diagnostic accuracy reviews. Syst Rev 2(82):1–6
27. Schünemann HJ, Oxman AD, Brozek J, Glasziou P, Jaeschke R, Vist GE, Williams JW Jr, Kunz R, Craig J, Montori VM, Bossuyt P, Guyatt GH, GRADE Working Group (2008) Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. BMJ 336(7653):1106–1110
28. Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, Reitsma JB et al (2011) QUADAS-2: A revised tool for the quality assessment of diagnostic accuracy studies. Ann Int Med 155:529–536

29. Clifford T, Cohen JF, Deeks JJ, Gatsonis C, Hooft L, Hunt HA, Hyde CJ, Korevaar DA, Leeflang MMG, Macaskill P, Reitsma JB, Rodin R, Rutjes AWS, Salameh JP, Stevens A, Takwoingi Y, Tonelli M, Weeks L, Whiting P, Willis BH, the PRISMA-DTA Group (2018) Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies: the PRISMA-DTA statement. JAMA 319(4):388–396
30. Salameh JP, McInnes MDF, Moher D, Thombs BD, McGrath TA, Frank R, Sharifabadi AD, Kraajipoel N, Levis B, Bossuyt PM (2019) Completeness of reporting of systematic reviews of diagnostic test accuracy based on the PRISMA-DTA reporting guideline. Clin Chem 65(2):291–301
31. Nguyen AV, Blears EE, Ross E, Lall RR, Ortega-Barnett J (2018) Machine learning applications for the differentiation of primary central nervous system lymphoma from glioblastoma on imaging: a systematic review and meta-analysis. Neurosurg Focus 45:1–10
32. Ariji Y, Fukuda M, Kise Y, Nozawa M, Yanashita Y, Fujita H, Katsumata A, Ariji E (2018) Contrast-enhanced computed tomography image assessment of cervical lymph node metastasis in patients with oral cancer by using a deep learning system of artificial intelligence. Oral Surg Oral Med Oral Pathol Oral Radiol 000:1–6
33. Choi H-S, Choe JY, Kim H, Han JW, Chi YK, Kim K, Hong J, Kim T, Yoon S, Kim KW (2018) Deep learning based low-cost high-accuracy diagnostic framework for dementia using comprehensive neuropsychological assessment profiles. BMC Geriatr 18:234–245
34. Lotsch J, Hummel T (2019) A machine-learned analysis suggests non-redundant diagnostic information in olfactory subtests. IBRO Rep 6:64–73
35. Choi KJ, Jang JK, Lee SS, Sung YS, Shim WH, Kim HS, Yun J, Choi J-Y, Lee Y, Kang B-K, Kim JH, Kim SY, Yu ES (2018) Development and validation of a deep learning system for staging liver fibrosis by using contrast agent-enhanced CT images of the liver. Radiology 289:688–697
36. Chuchu N, Dinnes J, Takwoingi Y, Matin RN, Bayliss SE, Davenport C, Moreau JF, Bassett O, Godfrey K, O'Sullivan C, Walter FM, Motley R, Deeks JJ, Williams HC, Cochrane Skin Cancer Diagnostic Test Accuracy Group (2018) Teledermatology for diagnosing skin cancer in adults. Cochrane Database Syst Rev 12:CD013193
37. Ferrante di Ruffano L, Takwoingi Y, Dinnes J, Chuchu N, Bayliss SE, Davenport C, Matin RN, Godfrey K, O'Sullivan C, Gulati A, Chan SA, Durack A, O'Connell S, Gardiner MD, Bamber J, Deeks JJ, Williams HC, Cochrane Skin Cancer Diagnostic Test Accuracy Group (2018) Computer-assisted diagnosis techniques (dermoscopy and spectroscopy-based) for diagnosing skin cancer in adults. Cochrane Database Syst Rev 12:CD013186
38. Knottnerus JA, Muris JW (2003) Assessment of the accuracy of diagnostic tests: the cross-sectional study. J Clin Epidemiol 56:1118–1128
39. Owen RK, Cooper NJ, Quinn TJ, Lees R, Sutton AJ (2018) Network meta-analysis of diagnostic test accuracy studies identifies and ranks the optimal diagnostic tests and thresholds for health care policy and decision-making. J Clin Epidemiol 99:64–74
40. Kang SK (2019) Measuring the value of MRI: comparative effectiveness and outcomes research. J Magn Reson Imaging 49(7): e78–e84. https://doi.org/10.1002/jmri.26647
41. Ransohoff DF (2002) Challenges and opportunities in evaluating diagnostic tests. J Clin Epidiomol 55:1178–1182
42. Lord SJ, Irwig LE, Simes RJ (2006) When is measuring sensitivity and specificity sufficient to evaluate a diagnostic test, and when do we need randomized trials? Ann Int Med 144:850–855
43. Vos LM, Bruning AHL, Reitsma JB, Schuurman R, Riezebos-Brilman A, Hoepelman AIM, Oosterheert JJ (2019) Rapid molecular tests for , respiratory syncytial virus, and other respiratory viruses: a systematic review of diagnostic accuracy and clinical impact studies influenza. Clin Infect Dis 69(7):1243–1253. https://doi.org/10.1093/cid/ciz056
44. Rodger M, Ramsay T, Fergusson D (2012) Diagnostic randomized controlled trials: the final frontier. Trials 13(137):1–7
45. Lijmer J, Bossuyt PMM (2009) Various randomized designs can be used to evaluate medical tests. J Clin Epidemiol 62:364–373
46. di Ruffano LV, Hyde CJ, McCaffrey KJ, Bossuyt PMM, Deeks JJ (2012) Assessing the value of diagnostic tests: a framework for designing and evaluating trials. BMJ 344 (e686):1–9
47. Lu B, Gatsonis C (2012) Efficiency of study designs in diagnostic randomized clinical trials. Stat Med 32(9):1451–1466
48. Dowdy DW, Gounder CR, Corbett EL, Ngwira LG, Chaisson RE, Merritt MW (2012) The ethics of testing a test: randomized

trials of the health impact of diagnostic tests for infectious diseases. CID 55:1522–1526

49. Novielli N, Sutton AJ, Cooper NJ (2013) Mata-analysis of the accuracy of two diagnostic tests used in combination: Application to the Ddimer test and the Wells Score for the diagnosis of deep vein thrombosis. Value Health 16:619–628

50. Wang J (2019) Analysis of the application values of different combination schemes of liquid-based cytology and high-risk human papilloma virus test in the screening of high-grade cervical lesions. Braz J Med Biol Res 52 (1):e7517. https://doi.org/10.1590/1414-431X20187517