Clinical Microbiology and Infection 25 (2019) 54–59

Contents lists available at ScienceDirect

Clinical Microbiology and Infection

journal homepage: www.clinicalmicrobiologyandinfection.com

Review

How to: evaluate a diagnostic test

M.M.G. Leeflang [1], [*], F. Allerberger [2]

[1)] *Department of Clinical Epidemiology and Biostatistics, University of Amsterdam, Amsterdam, The Netherlands*
[2)] *Division of Public Health, Austrian Agency for Health and Food Safety, Vienna, Austria*

ARTICLE INFO

Article history:
Received 22 January 2018
Received in revised form
1 June 2018
Accepted 2 June 2018
Available online 12 June 2018

Editor: L. Leibovici

Keywords:
Clinical epidemiology
Diagnostic research
Diagnostic test accuracy
In vitro diagnostics
Medical test evaluation
Study designs

ABSTRACT

Background: The development of an *in vitro* diagnostic test from a good idea to a clinically relevant tool takes several steps, with more stringent requirements at every step.
Objectives: This article aims to summarize the necessary questions to be asked about a test and to illustrate study designs answering these questions. We also aim to relate Regulation (EU) 2017/746 to the needs of evidence-based diagnostic testing, where applicable.
Sources: We used literature on evidence-based diagnostics, a text book on clinical trials in the development and marketing of medical devices and the English version of Regulation 2017/746 of the European Parliament and of the Council on *in vitro* diagnostic medical devices.
Content: The combination of different test uses and different stages of development determine the required test characteristics and suitability of study designs. In an earlier stage of test development it may be crucial to know whether a test can differentiate diseased persons from healthy controls, although this tells us little about how a test will perform in practice. Later stages focus on the diagnostic accuracy of a test in a clinically relevant situation. However, a test that perfectly distinguishes between patients with and without a certain condition may still have little effect on patient outcomes. Therefore, randomized controlled trials of testing may be needed, as well as post-marketing monitoring.
Implications: Both researchers and users of tests need to be aware of the limitations of diagnostic test accuracy and realize that accuracy is only indirectly linked to people's health status. **M.M.G. Leeflang, Clin Microbiol Infect 2019;25:54**

Background

Any intervention, medical device or test used in health care should eventually be beneficial for the persons subjected to them. Using a diagnostic test just for the sake of testing may lead to contradictory and therefore confusing test results, to overdiagnosis and to unnecessary burden and costs. Although a diagnostic test may be expected to indicate accurately the risk of having a certain condition, testing should ultimately lead to an improvement of the health status of the tested person [1].

In the area of infectious diseases, most clinical tests used to arrive at a diagnosis will be so-called *in vitro* tests. These are medical devices for the examination of specimens derived from the human body (such as blood, sputum, tissue) for the purpose of providing information on the current or future status of the person tested [2]. These include risk profiling, genetic information and prediction of treatment response.

On 5 April 2017, the European Parliament issued a new Regulation (EU) 2017/746 on diagnostic medical devices. It aims to set high standards of quality and safety for *in vitro* diagnostic medical devices by ensuring, among other things, that data generated in performance studies are reliable and robust and that the safety of subjects participating in performance studies is protected [2]. Approval by a notified body will be more often needed than previously and the notified bodies will be put under strict EU peer review. The performance evaluation of a diagnostic medical device is a continuous process by which data are assessed and analysed to demonstrate the scientific validity, and analytical and clinical performance of the device for its intended purpose as stated by the manufacturer. The manufacturer shall establish and update a performance evaluation plan following a defined and methodologically sound procedure in order to demonstrate (a) scientific validity, (b) analytical performance and (c) clinical performance. The data and conclusions drawn from the assessment of those elements

[*] Corresponding author. M.M.G. Leeflang.
E-mail address: m.m.leeflang@amc.uva.nl (M.M.G. Leeflang).

https://doi.org/10.1016/j.cmi.2018.06.011
1198-743X/© 2018 European Society of Clinical Microbiology and Infectious Diseases. Published by Elsevier Ltd. All rights reserved.

constitute the clinical evidence for the device [2]. Although it is not yet clear what these studies exactly should be like to be 'methodologically sound' or how the results of the three levels of evaluation should be interpreted for clinical use, extra systems will be put in place to ensure that manufacturers employ transparency and accountability.

In most countries, including those of the European Union, diagnostic devices are differently regulated from drugs [3]. If improving a person's health status is the starting point, then the requirements for a medical test should not be that much different from those set for drugs and pharmaceuticals. And the steps to follow from development to clinical use may follow more or less the same phased approach: from early development and technical performance via diagnostic test accuracy to clinical impact and (cost-) effectiveness [4–6]. The EU Regulation seems to focus on the first three stages only, although it briefly addresses post-market performance follow up.

This 'How to' article describes the four main stages of test development and evaluation, indicating at every step the questions that need to be asked and the study designs that may best address these questions. We provide a detailed discussion of the features of diagnostic tests that can pose challenges for the design of well-controlled clinical studies as well as methods for addressing these design challenges. Our focus will be mainly on the use of *in vitro* diagnostic tests. We furthermore aim to relate the EU Regulation to the needs of evidence-based diagnostic testing, taking into account that the Regulation may be interpreted in different ways and that neither of the authors of this 'How to' article was involved in the Regulation.

Phase 1: discrimination between two groups in an ideal situation

First of all, there must be an association between the test and the condition of interest. In the EU Regulation, this is translated as 'scientific validity of an analyte' and defined as 'the association of an analyte with a clinical condition or a physiological state' [2]. This means that for a continuous test, the mean value of the test result should be different in patients with the disease of interest and in those without it. For a categorical or dichotomous test result, this means that the percentage of positive or negative test results should be different in both groups. According to the EU Regulation, this information may come from scientific literature, expert opinion, proof of concept studies or clinical performance studies [2].

An example of a study providing proof of concept for the relationship between four serological biomarkers and *Pneumocystis* pneumonia may be found in Esteves *et al.* [7]. This study included 260 participants, from different locations, either with or without *Pneumocystis* pneumonia. Forty blood donors with no proof of *Pneumocystis* pneumonia were added to the group without pneumonia. It may be expected that none of the evaluated biomarkers will turn out to be elevated in many of those healthy controls. If most of the controls turn out to be positive after testing, the initial concept of association between biomarker and disease cannot be proven. This study provided p-values to indicate the difference between the levels of biomarkers in the different groups, indicating a statistically significant difference between the cases and controls. However, the figures in the publication show much overlap in biomarker levels between the different participant groups.

Another example can be found in a study evaluating the accuracy and safety of a new skin-test reagent for detecting tuberculosis [8]. Here, a group of healthy controls was included with a group of patients with definitely diagnosed active tuberculosis. In this study, very low p-values accompanied a very high area under the receiver operating characteristic curve, indicating that the discriminatory ability of the skin test was very high. If in such a population the test had not had high accuracy, the concept would not have been proven.

At this stage, sensitivity is often estimated in the diseased patients and specificity in healthy samples. One could call this a two-gate design, with the diseased participants coming from a different location (through a different gate) than the healthy controls [9]. Often, the definition of the cases is based on the agreed reference standard, whereas the selection of the controls is based on different criteria. The results from such a design are often presented as if they show the sensitivity and specificity of a test. However, the assumption that these measures are indeed useful for clinical practice is problematic. First of all, if the cases have typically very extreme values of what needs to be measured, then the test will be positive in most of the patients. Similarly, if the volunteers are healthy and have test results at the lower extreme end of the testing spectrum, then the test will be negative in most of those cases. This calls into question the applicability of the study results for clinical practice. In clinical practice, some patients without the disease may have co-morbidities resulting in test results similar to those of the patients with the disease, causing a lower specificity. On the other hand, some patients with the disease will present as less extreme in clinical practice, resulting in lower sensitivity.

Irrespective of their risk of biased results and limited applicability, these severe-cases versus healthy-control designs have their place in test evaluation research. Indeed, without proof-of-concept, there is no need to further develop the tests. If the test performs poorly in such a design (with and without the target condition), then we can be quite sure the test will perform at best equally poorly in a clinical setting.

Not all studies evaluating diseased and non-diseased participants separately should be set aside as two-gate designs. In some situations, for example low prevalence settings, it may be more efficient to sample all diseased persons and only a random subset of the non-diseased. In this case, the required sample size is much lower, while the controls are still representative of the non-diseased persons in clinical practice.

Phase 2: technical validity

If a test seems to be capable of distinguishing patients with the target condition from patients without the target condition, we need to make sure it is not a one-off result. Technical validity therefore covers a very broad range of questions and outcomes. It is about research into the repeatability and reproducibility of laboratory tests, about inter-rater research of imaging tests, but also about analytical sensitivity (minimally detectable levels).

In the EU Regulation, this is referred to as analytical performance, defined by outcome measures, such as 'analytical sensitivity, analytical specificity, trueness (bias), precision (repeatability and reproducibility), accuracy (resulting from trueness and precision), limits of detection and quantitation, measuring range, linearity, cut-off, including determination of appropriate criteria for specimen collection and handling and control of known relevant endogenous and exogenous interference and cross-reactions' [2].

Although in most presented phased approaches technical validity comes before all other phases, it may make more sense in practice to investigate technical validity once there is some evidence of association between test result and disease status. In publications, we often see that phase 2 is combined with phase 1 or phase 3; the samples and data of a phase 1 or phase 3 study can also be used for a phase 2 study or vice versa. Here again, the exact analyses and outcome measures may depend on the test type and whether the test results are continuous, categorical or dichotomous.

Continuous test results

Analytical variability may impact the accuracy of a test (calibration) and its precision (consistency, reproducibility). For continuous test results it is important that the measurement be stable: if it is repeated several times, does it produce the same result? This can be investigated using Bland–Altman plots, in which the average of two measurements is plotted against the difference between two measurements [10]. This may be done on the same sample and under the same circumstances, which is often referred to as repeatability. But it may also be done between assessors, between machines or between laboratories. In this case it is called reproducibility. If the results differ, this may be because of differences in the way the test is being performed; but it may also be because the extent of biological variation is large.

Categorical or dichotomous test results

For dichotomous or categorical tests too, repeatability and reproducibility are important. Then it is often referred to as interrater reliability or interobserver reliability. This can be expressed using κ statistics, which quantify the agreement between observers or measurements beyond chance alone [11]. However, just crude agreement is often easier to interpret and to estimate [12].

The multicentre evaluation of antifungal susceptibility testing methods of the European Committee on Antimicrobial Susceptibility testing is an example of a reproducibility study, both for categorical outcomes and continuous outcomes [13].

Phase 3: clinical validity and accuracy in a clinically relevant situation

If the test seems to be able to distinguish patients with the target condition from those without, and if the test results seem to be sufficiently robust, then it is time to move from laboratory samples and extreme participant samples to a real life situation in a prospectively planned and conducted study. In real life, the patients to be tested will be symptomatic, unless the test is intended as a screening test. Also, the samples may not be as perfect as in a laboratory situation. Although this may also be true for the technical validity of a test (reproducibility of blood pressure results may, for example, be different in healthy men and in healthy pregnant women) we focus here on the accuracy of a diagnostic test.

The EU Regulation refers to this stage as the clinical performance of the device and states that the manufacturer should demonstrate this ‘in relation to diagnostic sensitivity, diagnostic specificity, positive predictive value, negative predictive value, likelihood ratio and expected values in normal and affected populations’. It does not make a distinction between laboratory settings and real-life situations, but it also states that devices should be suitable for their intended purpose [2].

Who will be tested?

The first question to be answered is who will be tested with the test under evaluation in real life. We need to know where these patients come from, through what referral route, in what healthcare setting they will be tested and what the management steps after a negative or a positive test result will be. A test for leptospirosis may require different performance characteristics if it is the first test in line, used to triage who will be tested further than when the test is used to confirm infection. Also, the same test may have different performance characteristics in a triage situation than in a confirmation setting. In a confirmation setting, most patients who clearly do not have the disease of interest will be excluded [14].

What to explain about the tests of interest?

Only providing the sensitivity and specificity (or predictive values) of the test of interest may not be helpful for clinical decision making. For example, if a new test may be used to replace an older test, you need to know how the new test compares to the older test. Is it more accurate? Or is it equally accurate combined with better feasibility or lower costs? Is there a choice between multiple tests at the same time? In that case, you may need to compare all tests available. If the test is going to be used as a triage test, then you may want to compare the accuracy of the current testing strategy with the testing strategy that includes the new test as a triage test [15].

The way the tests were conducted is important as well. For example, the training and expertise of the microscopist influences the number of false positives and false negatives after microscopy. Also, the time it takes to look through the microscope may be relevant and should be reported. For more subjective testing, the amount of additional information provided to the assessor may be relevant, especially if the assessor knew whether the sample came from a person with or without disease.

An integral part of assessing the sensitivity and specificity of a continuous diagnostic test is the threshold at which it is to be used. If the matter that is to be measured, for example antibody levels, is higher in patients who are more likely to be diseased, a higher positivity threshold will result in lower sensitivity and higher specificity. However, the same positivity threshold may not result in the same sensitivity and specificity, especially not in the case of serological tests for infectious diseases. In a high-prevalence setting, more people may have higher antibody levels without being diseased, resulting in a lower specificity. Choosing the optimal threshold may be done in different ways, but one has to realize that determining an optimal cut-off point based on the data at hand may lead to overoptimistic and therefore biased results [16].

What is the target condition?

The target condition is a specific manifestation of the disease that the test is intended to detect. For example, if the disease is tuberculosis, then the target condition that we are specifically interested in, may be extra-pulmonary tuberculosis, or multidrug-resistant tuberculosis, or tuberculosis that is responsive to treatment. Each of these target conditions may require a different reference standard to assess the true status of the patient. The reference standard used to be called the gold standard. For most diseases there is no gold standard and if there is one, then it may not reflect the clinical decision that needs to be made for this patient. The reference standard is sufficiently reliable to select those patients, e.g. the patients that really require treatment, whereas the gold standard may detect everyone with the faintest sign of disease, which may even lead to over-diagnosis and unnecessary treatment.

Although we claim that a gold standard seldom exists and is rarely clinically relevant, we must have some faith in the idea that the reference standard is sufficiently reliable to differentiate between people with or without the target condition. If that is not possible either, then other solutions need to be found [17]. The easiest solution is to take the imperfectness of the reference standard into account when interpreting the results. For example, a sensitivity of 80% does not mean that 80% of people with malaria will be detected, but that the rapid test is positive in 80% of the cases with a positive blood smear. Other solutions may be the use of composite reference standards: use all tests or clinical criteria available for a diagnosis. However, if the test under evaluation is part of this composite reference standard, its results will influence the final diagnosis, which may lead to a bias. Expert opinions may also be used to determine who has the target condition, but they are not always reliable. Statistical solutions include latent class analysis: a statistical technique in which a model estimates the chance of a patient having the target condition or not.

The clinical pathway

The choices to be made in designing an accuracy study may be outlined by drawing a clinical pathway. This is a graphical representation of the diagnostic path that the patient follows from initial suspicion to final treatment. It may show that some tests will not be used in some places and therefore will not be competitors for the test of interest. But these tests may serve as valid reference standards, as they may be used satisfactorily to guide treatment or further testing. Furthermore, the clinical pathway may guide researchers in the questions of whether multiple tests are involved and whether these multiple tests can be used in parallel or serially. It also helps to understand the consequences of using tests in parallel or serially.

The clinical pathway also helps to understand the consequences of true and false positive or negative test results. Will patients with a positive test result be referred for further testing, or will they be treated without further testing? These different choices may have a different impact on the individuals with false-positive results: if treatment is very burdensome, then a false-positive result may be even less desirable than in a situation where the false-positive result will be checked first by another test or device.

Study design and data analysis

The ideal diagnostic accuracy study starts by including all patients who will be tested in practice with the test(s) of interest. Patients fulfilling the inclusion criteria should be enrolled consecutively, without any judgement about how likely this person is to become test positive or test negative (Fig. 1). Also, patients who can be expected to have false-positive or false-negative test results but who will be tested in practice as well, should not be excluded to make the test appear better.

Hence, all these patients should undergo the test(s) of interest and its competitors, if relevant. So if, for example, the question is whether rapid test A or rapid test B should be used to detect malaria in children, they should both be applied to all children who should be tested. The assessment of these tests should be done without knowing the results of any of the other tests, including the reference standard. A post-hoc selection of the optimal cut-off point for test positivity may lead to overoptimistic sensitivity and specificity. The cut-off point at which any continuous test is used should be specified in the study protocol in advance. Usually this will be the cut-off value as recommended by the manufacturer.

All included patients should undergo the reference standard test and the assessment of this test should be done without knowing the results of the tests under evaluation. If the reference standard test is costly or the prevalence of the target condition is too low, then a random selection of test negatives or test positives may be required rather than testing all included patients.

The data may be presented in two-by-two tables from which sensitivity, specificity and all other accuracy measures can be estimated (Table 1). Although for test manufacturers sensitivity and specificity may be important, as they indicate the ability of a test to indicate who has the disease and who has not, for a clinician taking care of an individual patients, the predictive values may be more relevant. These predictive values are directly dependent on prevalence and therefore even more setting-specific than sensitivity and specificity. In studies estimating sensitivity and specificity separately in diseased and non-diseased (case—control designs, two-gate designs), the researcher determines the ratio between diseased and non-diseased (= prevalence) and therefore the resulting predictive values make no sense. However, in the case of multiple tests, it may be useful to present all test result combinations and the numbers of patients per combination. From such a table, a 2 × 2 table for each test can be derived, but also the correlations between tests can be inferred and the accuracy of test combinations can be deduced [18].

As the performance of a test may differ between situations, reporting the study characteristics is crucial for interpretation of the study results and conclusions. The Standards for Reporting of Diagnostic accuracy studies may be helpful for authors when writing their report [19].

Phase 4: effect on outcomes important for the patient and on society

The accuracy of a test says little about whether patients who undergo the evaluated test(s) indeed fare better than similar patients who have not been tested or who have been tested with a different test. Therefore, in the next stage researchers investigate to

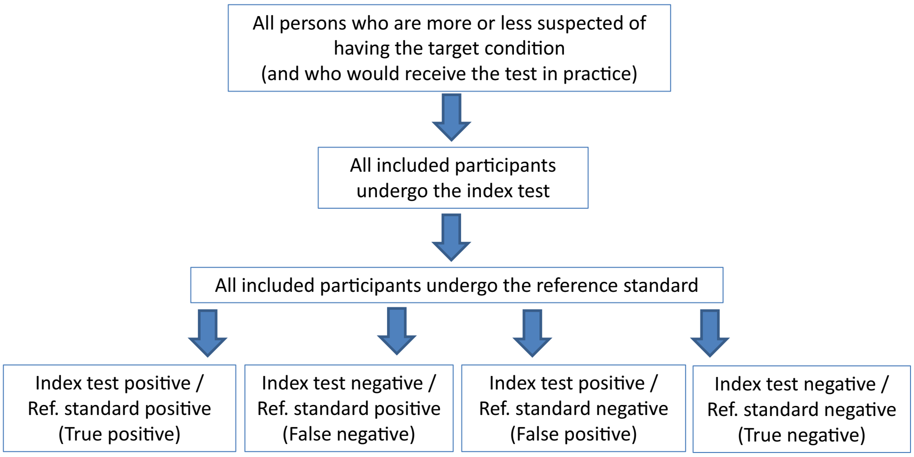
All persons who are more or less suspected of having the target condition (and who would receive the test in practice)
All included participants undergo the index test
All included participants undergo the reference standard
Index test positive / Ref. standard positive (True positive)
Index test negative / Ref. standard positive (False negative)
Index test positive / Ref. standard negative (False positive)
Index test negative / Ref. standard negative (True negative)

Fig. 1. Flow diagram of a test evaluation study, leading to an estimate of clinical accuracy of the investigated test.

Table 1
Two-by-two table and definitions for measures of test accuracy [2]

	Reference standard positive (Disease present)	Reference standard negative (Disease absent)	
Index test positive	a	b	a+b
Index test negative	c	d	c+d
	a+c	b+d	

Diagnostic sensitivity	The ability of a device to identify the presence of a target marker associated with a particular condition, expressed as the percentage of persons with the target condition who test positive (a/(a+c)).
Diagnostic specificity	The ability of a device to recognize the absence of a target marker associated with a particular condition, expressed as the percentage of persons without the target condition who test negative (d/(b+d))
Positive predictive value	The ability of a device to separate true-positive results from false-positive results for a given attribute in a given population, expressed as the percentage of persons with a positive test result who indeed have the disease (a/(a+b))
Negative predictive value	The ability of a device to separate true-negative results from false-negative results for a given attribute in a given population, expressed as the percentage of persons with a negative test result who do not have the disease (d/(c+d))
Likelihood ratio	The likelihood of a given result arising in an individual with the target clinical condition compared with the likelihood of the same result arising in an individual without that clinical condition.

what extent the test results change diagnostic reasoning and decision making. But the ultimate question should be: ‘Does the patient's situation improve after testing?’ So the relevant outcomes should also reflect health outcomes that are important for the patient. Whether a test changes the decision-making process, changes the choice of treatment, or enhances the confidence of the doctor treating the patient is only relevant if the patient indeed benefits from testing. Patient-relevant outcomes may be risk of disease, risk of death, or patient's quality of life.

The ultimate study design to assess the effect of a test on patients' health is the randomized controlled trial. In such a randomized controlled trial, one patient group will undergo the test of interest and will be further managed according to the test results, while the other patient group will be managed in a different way. This different way may be an alternative test, or referral based on clinical signs and symptoms, or treatment based on a different algorithm than the test under evaluation. For example, a study performed in Ghana included all patients visiting health-care facilities and for whom the consulted health-care professionals considered malaria and wanted to test for malaria or treat the patient with an antimalarial [20]. Among other outcomes, the study showed that in settings without access to microscopy the use of rapid diagnostic tests led to a significant reduction in the prescription of antimalarials. This outcome may be a surrogate outcome for patient-relevant outcomes, as it indicates that patients receiving the right treatment will fare better, while not directly assessing the status of the patients.

Such a design has drawbacks. The expected difference in outcomes between the two groups will typically be small, and so such a randomized controlled trial requires a large sample size for sufficient power to detect existing differences. Furthermore, the protocol for both arms needs to be clear and has to be followed correctly.

The last step to consider may be cost-effectiveness and the potential effects on society of introducing the test. This may be estimated through modelling and in real-time can be monitored by post-market surveillance and performance follow up. Under the new Regulation, manufacturers have to produce periodic safety and summary reports, summarizing the results and conclusions of post-marketing surveillance. They also have to take corrective actions in case of potentially serious incidents [2].

Discussion

Although the benefit of diagnostic medical devices lies in providing accurate medical information about patients, the concept of clinical benefit from diagnostic medical devices is not fundamentally different from the concept of clinical benefit from pharmaceuticals or therapeutic medical devices. In the end, anything we apply to patients and to the public should be beneficial for them and should do no harm nor lead to unnecessary burden. Yet, the final clinical outcome for the patient does not depend on the test under evaluation only, but is dependent on further diagnostic and therapeutic options. Although it is clear that tests need to pass through several stages before we can agree on their value for patients or society, in practice most tests are being used without such a rigorous evaluation plan.

As a general rule, devices should bear the Conformitée Européenne (CE) marking to indicate their conformity with Regulation (EU) 2017/746 [2]. However, this CE mark says nothing about the clinical performance of a test. For example, the performance of a CE-marked home test for *Chlamydia trachomatis* available on the internet was found to be very poor. The test yielded more false-positive results (18/193) than true-positive results (7/38) [21]. Still, this test remains accessible via the internet under various brand names and has retained its CE mark. This situation raised serious concerns about the regulation of diagnostic products available via the internet and the standards of certain Notified Bodies that issue the CE mark [21]. Under the new Regulation, these situations should no longer occur.

In the near future, the traceability of devices by means of the Unique Device Identification system (UDI system) will enhance the effectiveness of post-marked safety-related measures for devices. However, whether the new Regulation and this UDI leads to marketing of tests that are more suitable for their intended purpose remains to be seen. Focusing only on the clinical accuracy of a test may not be sufficient to indicate whether the use of a test is beneficial for patients and for society.

In conclusion, both researchers and users of tests need to be aware of the limitations of diagnostic test accuracy and realize that accuracy is only one dimension of the value or usefulness of a diagnostic test. Accuracy is an indirect predictor of the effect of the test on people's health status. When designing or reading an accuracy study, people should realize that test accuracy, including sensitivity and specificity, may vary in different situations. In any case, healthy controls are usually not representative for persons tested in practice and should no longer be included in accuracy studies that claim to be clinically relevant.

Transparency declaration

Dr Leeflang and Prof. Allerberger have nothing to disclose. No external funding was received for any part of this manuscript. Dr Leeflang wrote the first draft, both authors wrote parts of subsequent versions and agreed on the final version.

References

[1] Ferrante di Ruffano L, Hyde CJ, McCaffery KJ, Bossuyt PM, Deeks JJ. Assessing the value of diagnostic tests: a framework for designing and evaluating trials. BMJ 2012;344:e686.

[2] Regulation (EU) 2017/746 of the European Parliament and of the Council of 5 April 2017 on *in vitro* diagnostic medical devices and repealing directive 98/79/EC and Commission Decision 2010/227/EU. Offic J Eur Union 2017;L 117: 176–331.

[3] Becker KM. Clinical trials in development and marketing of medical devices. In: Becker KM, Whyte J, editors. Clinical evaluation of medical devices. Principles and case studies. 2nd ed. Totowa, NJ: Humana Press; 2006. p. 3–19.

[4] Sackett DL, Haynes RB. The architecture of diagnostic research. BMJ 2002;324: 539–41.

[5] Guyatt GH, Tugwell PX, Feeny DH, Haynes RB, Drummond M. A framework for clinical evaluation of diagnostic technologies. Can Med Assoc J 1986;134: 587–94.

[6] Lijmer JG, Leeflang M, Bossuyt PM. Proposals for a phased evaluation of medical tests. Med Decis Making 2009;29:E13–21.

[7] Esteves F, Calé SS, Badura R, de Boer MG, Maltez F, Calderón EJ, et al. Diagnosis of *Pneumocystis pneumonia*: evaluation of four serologic biomarkers. Clin Microbiol Infect 2015;21:379.e1–10.

[8] Li F, Xu M, Qin C, Xia L, Xiong Y, Xi X, et al. Recombinant fusion ESAT6-CFP10 immunogen as a skin test reagent for tuberculosis diagnosis: an open-label, randomized, two-centre phase 2a clinical trial. Clin Microbiol Infect 2016;22. 889.e9–889.e16.

[9] Rutjes AW, Reitsma JB, Vandenbroucke JP, Glas AS, Bossuyt PM. Case–control and two-gate designs in diagnostic accuracy studies. Clin Chem 2005;51: 1335–41.

[10] Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;1:307–10.

[11] Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960;20:37–46.

[12] De Vet HC, Mokkink LB, Terwee CB, Hoekstra OS, Knol DL. Clinicians are right not to like Cohen's κ. BMJ 2013;346:f2125.

[13] Cuenca-Estrella M, Arendrup MC, Chryssanthou E, Dannaoui E, Lass-Flörl C, Sandven P, et al., AFST Subcommittee of EUCAST. Multicentre determination of quality control strains and quality control ranges for antifungal susceptibility testing of yeasts and filamentous fungi using the methods of the Antifungal Susceptibility Testing Subcommittee of the European Committee on Antimicrobial Susceptibility Testing (AFST-EUCAST). Clin Microbiol Infect 2007;13:1018–22.

[14] Mulherin SA, Miller WC. Spectrum bias or spectrum effect? Subgroup variation in diagnostic test evaluation. Ann Intern Med 2002;137:598–602.

[15] Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: assessing new tests against existing diagnostic pathways. BMJ 2006;332:1089–92.

[16] Leeflang MM, Moons KG, Reitsma JB, Zwinderman AH. Bias in sensitivity and specificity caused by data-driven selection of optimal cutoff values: mechanisms, magnitude, and solutions. Clin Chem 2008;54:729–37.

[17] 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.

[18] Hayen A, Macaskill P, Irwig L, Bossuyt P. Appropriate statistical methods are required to assess diagnostic tests for replacement, add-on, and triage. J Clin Epidemiol 2010;63:883–91.

[19] Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, et al., STARD Group. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. BMJ 2015;351:h5527.

[20] Ansah EK, Narh-Bana S, Epokor M, Akanpigbiam S, Quartey AA, Gyapong J, et al. Rapid testing for malaria in settings where microscopy is available and peripheral clinics where only presumptive treatment is available: a randomised controlled trial in Ghana. BMJ 2010;340:c930.

[21] Michel CE, Saison FG, Joshi H, Mahilum-Tapay LM, Lee HH. Pitfalls of internet-accessible diagnostic tests: inadequate performance of a CE-marked *Chlamydia* test form home use. Sex Transm Infect 2009;85:187–9.