Clinical Microbiology and Infection 30 (2024) 1261–1269

Contents lists available at ScienceDirect

Clinical Microbiology and Infection

journal homepage: www.clinicalmicrobiologyandinfection.com

Narrative review

How to verify and validate a clinical microbiology test before it can be used in routine diagnostics: a practical guide

Erlangga Yusuf [1,*], Maarten J. Schijffelen [2], Mariska Leeflang [3]

[1] Department of Medical Microbiology and Infectious Diseases, Erasmus University Medical Center, Rotterdam, the Netherlands
[2] LabMicta, Hengelo, the Netherlands
[3] Department of Epidemiology and Data Science, Amsterdam Public Health, Amsterdam UMC, Amsterdam, the Netherlands

ARTICLE INFO

Article history:
Received 27 February 2024
Received in revised form
29 June 2024
Accepted 29 June 2024
Available online 6 July 2024

Editor: L. Leibovici

Keywords:
Antimicrobial susceptibility testing
Diagnostic tests
ISO
Validation
Verification

ABSTRACT

Background: Before a new test can be routinely used in your laboratory, its reliability must be established in the laboratory where it will be used. International standards demand validation and verification procedures for new tests. The International Organization for Standardization (ISO) 15189 was recently updated, and the European Commission's *In Vitro* Diagnostic Regulation (IVDR) came into effect. These events will likely increase the need for validation and verification procedures.
Objectives: This paper aims to provide practical guidance in validating or verifying microbiology tests, including antimicrobial susceptibility tests in a clinical microbiology laboratory.
Sources: It summarizes and interprets certain parts of standards such as ISO 15189:2022, and regulations, such as IVDR 2017/746 regarding validation or verification of a new test in a routine clinical microbiology laboratory.
Content: The reasons for choosing a new test and the outline of the validation and verification plan are discussed. Furthermore, the following topics are touched upon: the choice of reference standard, number of samples, testing procedures, how to solve the discrepancies between results from new test and reference standard, and acceptance criteria. Arguments for selecting certain parameters (such as reference standard and sample size) and examples are given.
Implications: With the expected increase in validation and verification procedures because of the implementation of IVDR, this paper may aid in planning and executing these procedures.
Erlangga Yusuf, Clin Microbiol Infect 2024;30:1261

Introduction

On 26 May 2022, the European Commission's *In Vitro* Diagnostic Regulation (IVDR) [1], came into effect. IVDR sets requirements for *in vitro* diagnostic tests, classifying them from low to high risk into classes A–D. Except for Class A (tests for general laboratory use), all classes need Conformité Européenne (CE) certification by notified bodies. It also requires laboratory-developed tests (LDTs) without CE-marking to be validated according to the International Organization for Standardization (ISO) 15189:2022 [2] (that updates ISO 15189:2012), whereas previously this was not required. IVDR may thus increase the number of validation and verification procedures (see Table 1 for definitions).

This paper aims to provide practical guidance for validating or verifying microbiology tests in a clinical microbiology laboratory. It does so by summarizing and interpreting standards and regulations, offering arguments for selecting certain parameters, and providing examples. Identification of bacteria (e.g. microscopy and matrix assisted laser desorption/ionisation-time of flight (MALDI-TOF)) and molecular tests are outside the scope of this paper, but several principles can still be applied.

When is a new test required?

Laboratories endeavour to implement new tests based on manufacturers' claims of improved accuracy, faster results, lower costs, less invasive, less labour intensive, and easier interpretation of the results than existing tests. In addition, new bacterial

* Corresponding author: Erlangga Yusuf, Department of Medical Microbiology and Infectious Diseases, Erasmus University Medical Center, Rotterdam, the Netherlands.
E-mail address: angga.yusuf@gmail.com (E. Yusuf).

https://doi.org/10.1016/j.cmi.2024.06.028
1198-743X/

Table 1
Verification or validation (between brackets are the source of the statement)

	Verification	Validation
Definition	Confirmation of truthfulness, by which the laboratory confirms the established performance claims of a test (ISO 15189:2022) [2].	Confirmation of plausibility of the test for a specific intended use. The confirmation should be shown by providing objective evidence in the form of performance specifications that fulfil certain requirements (ISO 15189:2022) [2].
Aim	To verify the performance characteristics that have been reported by manufacturers of a test before it is implemented in the laboratory that is going to use it.	To demonstrate that the test performs accurately and reliably as intended.
Which test?	CE-IVD labelled [2], or a test that has been validated in other ISO 15189 accredited laboratories (own interpretation).	- CE-IVD labelled or a validated test that is used outside the originally intended scope, or has been modified (ISO 15189:2022) [2]. - Laboratory-developed test (ISO 15189:2022) [2]. - Research Use Only labelled test that is going to be used in the diagnostic (own interpretation).
Frequency	The verification procedure should be performed only once as long as nothing is changed.	The validation procedure itself should be performed only once, but an ongoing process is needed to monitor whether the test is still fit for the intended purpose by participation in the quality assessment plan [2,28].
Reference standard	Existing tests in the laboratory or reference tests [6]. Be cautious that using different types of tests as the reference standard (e.g. T-spot test vs. tuberculin skin test) may affect the performance specifications of the new test. This should be taken into account when defining acceptance criteria. Clinical criteria can be used but they should fulfil the conditions needed for the clinical reference standard.	Ideally, the reference standard should include clinical or diagnostic criteria since validation aims to demonstrate that the test fits its specific intended use, which is clinical (either for screening, diagnostic, or confirmation) of a disease. Moreover, the validation should be 'as extensive as necessary' (ISO 15189:2022) [2]. This can be done when the new test is prospectively run in parallel in a certain time period in certain patients' populations to reflect prevalence in the specific patient population so clinical sensitivity and specificity can be calculated. However, in practice, other comparable tests are frequently used as reference standards.
Performance specifications (parameters)	Accuracy, precision (reproducibility), and reportable range [2].	Accuracy, precision (reproducibility), analytical sensitivity (i.e. limit of detection and limit of quantification), analytical specificity (e.g. cross-reaction with interfering substances), diagnostic sensitivity, and diagnostic specificity [2]. For quantitative tests, also linearity and reportable range.

ISO, International Organization for Standardization; CE-IVD.

resistance mechanisms and new antibiotics necessitate updated antimicrobial susceptibility tests (ASTs). Sometimes, manufacturers introduce new ingredients for culture media and AST's formulation.

What sort of tests?

Quantitative tests provide continuous results, such as serology tests measuring IgG and IgM concentrations. These results are often categorized as positive or negative using cut-offs (semi-quantitative). Qualitative tests only show positive or negative results, such as growth in culture. It is important to determine a test's position in the diagnostic cascade (Fig. 1) as performance requirements vary [3,4]. For example, tests for safely discharging patients need high sensitivity (low false negatives), whereas tests for conditions requiring long-term antimicrobial treatment (e.g. invasive aspergillosis) require high specificity (low false positives). The screening test and the diagnostic test may be the same but used in different ways. At the laboratory level, there are screening tests that need confirmation, such as chromogenic agars confirmed by a polymerase chain reaction (PCR) test to detect vancomycin-resistant enterococci [5].

Differences between validation and verification

Validation is a process to show that a test works as intended [2]. It is required for non-CE-IVD marked (usually LDTs), modified CE-IVD tests, or when planning to use CE-IVD tests in different specimen types than intended (e.g. using *Legionella* antigen test in pleural fluid instead of the intended use in urine). According to ISO

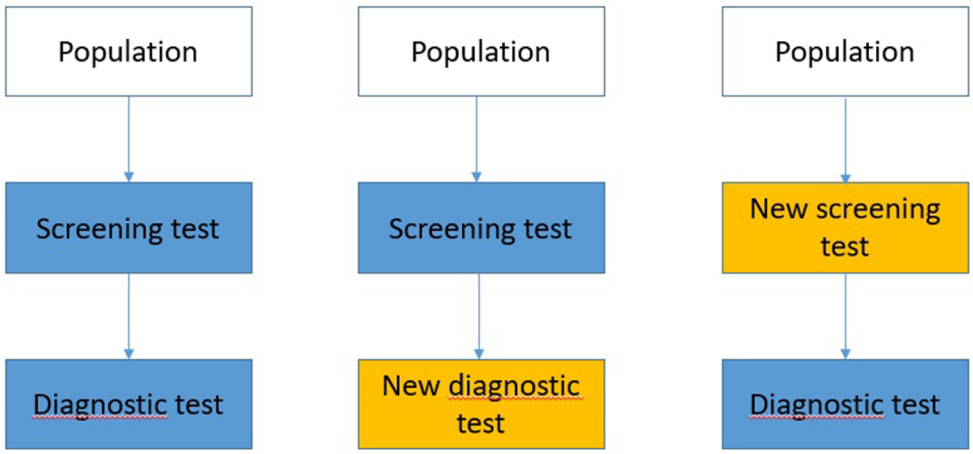

Fig. 1. Possible position of a test in diagnostic cascade.

15189, validation should be as extensive as necessary and confirm that the test fits its intended use. In our opinion, this requires correlation with clinical findings. For example, validating a new *Aspergillus* antigen assay involves testing its performance against the diagnosis of invasive pulmonary aspergillosis. Only then, clinical sensitivity and clinical specificity can be reported. These parameters depend on several factors such as when the test is performed during the course of the disease, and the patient's population (e.g. tests' sensitivity may differ between neutropenic and non-neutropenic patients) and they should be considered when planning a validation procedure. In practice, clinical data are often unavailable, or too labour intensive for laboratories to obtain. Instead, correlation with other non-clinical reference standards is used (see below), preventing the calculation of clinical sensitivity and specificity.

In a validation procedure, comprehensive parameters should be reported (Table 2) [2]. IVDR requires post-market surveillance for ongoing safety and efficacy, applicable for both CE-IVD tests and LDTs. After validation and implementation, periodic monitoring through internal and external quality control programme is required [1]. Frequent abnormal results in routine use, require re-evaluation of test performance.

Verification is aimed at verifying the performance characteristics reported by the manufacturers under the condition that the test is used as described by them. For example, an *Aspergillus* antigen test is intended for bronchoalveolar lavage fluid and must be verified in bronchoalveolar lavage fluid. Using it in cerebrospinal fluid requires validation. Verification is needed, because transportation, storage, handling of the test by technicians, and local conditions can affect test performance. Verification is a one-time procedure unless there are changes in test ingredients, sample types, or the population where it is performed. As verification aims at confirming the performance of a test as mentioned in the package insert, accuracy and precision are sufficient parameters in a verification procedure. ISO 15189 also lists reportable range as a verifiable parameter [2], but finding high-concentration samples can be challenging for the laboratory. Determining a lower limit of detection is labour intensive, involving spiking samples and performing data and statistical analysis. In our opinion, it is often unnecessary as semi-quantitative tests typically categorize low results as negative.

Validation/verification plan

A verification plan should be written before starting practical work, and no changes should be made once the verification begins. The plan (see also Table 4) should document the background and purpose of the evaluation, the reference standard, the number and which specimen panel to be tested, and the performance and acceptance criteria for the test to be used in the laboratory. It should include administrative information such as personnel involved in the verification procedure, target completion date, referral to relevant standard operating procedures, and references.

Choosing a reference standard

The term ‘reference standard’ is a more precise term than ‘gold standard’ for validation or verification [7,8] as no perfect gold standard exists, and pathognomonic signs for certain diseases may be rare or unspecific. Validation can establish a new test's clinical sensitivity and specificity, demonstrating its ability to discriminate between patients with and without a certain condition. A target condition (‘disease’) as a reference standard can be straightforward, e.g. septic arthritis with bacteria in a joint. Sometimes, a ‘disease’ is defined by the combination of signs, symptoms, and laboratory and radiology results. For example, the case definition of probable invasive aspergillosis needs a host factor (e.g. neutropenia), and radiological and microbiological findings [9]. Missing data on any criterion may cause misclassification [10]. If radiology is not performed, a patient cannot be categorized as having probable invasive aspergillosis [9]. In addition, a criterion (i.e. radiology) may carry more weight than other criteria, skewing comparisons to this heavy-weighted criterion, rather than the complete clinical criteria set.

In the verification procedures, the current test often serves as the reference to compare the new test [4]. This comparison may show that the new test provides the same or similar results as the old one, but it does not determine the true sensitivity and specificity of the new test for detecting a disease. Usually, the new test and the old test are (rather) of comparable types, e.g. serology using Enzyme Linked Immunosorbent Assay (ELISA) and Chemiluminescence Immunoassay. Using different reference standards affects the performance of the new test. For example, a new *Aspergillus* antigen test may be less sensitive compared with PCR, but more sensitive than fungal culture. If the new test is more sensitive than the reference standard, the reference standard may mislabel true positives as false positives, lowering its perceived specificity.

Number of samples

Validation requires more samples than verification because of the requirement of a thorough investigation of the new test (Table 3). Sample size calculation depends on the desired precision of sensitivity and specificity estimates or a pre-defined minimal estimate [11,12]. Smaller samples will result in less precise accuracy estimates (Fig. 2) and if the new test fails to detect only one out of 10 positive samples, sensitivity drops from 100 to 90% (95% CI: 55.5–99.8%).

Testing procedures

Prospective in parallel

For a test of prevalent diseases (e.g. urinary tract infection), a specific time period (e.g. 2 weeks) can be planned to collect the required number of samples. During this period, the new and reference tests can be simultaneously tested on patients' specimens from the population where the test will be used (i.e. general practitioners' patients), prospectively.

Testing in parallel

When disease prevalence is low, collecting enough samples for prospective testing of new and reference tests can be time-consuming. Instead, remnant samples that have been collected in the laboratory, can be tested by new and reference tests in parallel, though the contamination risk and deterioration from freeze and thaw cycles.

Alternatively, materials can be spiked with positive isolates, or quality control strains (e.g. American Type Culture Collection (ATCC) strains). The type of material (matrix) that is spiked, should match the patient specimen (e.g. *Histoplasma* antigen should be spiked in the urine and not in broth). Clearly, when selected clinical samples are used, the clinical performance of the test cannot be given. For verification, it is recommended that most positive samples should reflect those commonly found in the laboratory, and to include as many variations as possible. For example, validating a new IgM test for *Leptospira interrogans* in the Netherlands should mainly include serovars Icterohaemorrhagiae and Copenhageni, supplemented with other common serovars. For verification, both weakly and strongly positive specimens should be included to verify the reportable range of the test [6] and a limited number of ‘difficult’ samples, i.e. samples with the concentration around the

Table 2
Definitions of performance characteristics of a new test

	Definition	How to measure and calculate	Remarks and explanation
Accuracy	The proportion of new tests that give correct results as the reference standard (both positive and negative).	(Correct positive + correct negative)/total number [29].	- The most important parameter that determines the acceptance of a new test to be used in validation and verification [2]. - Applicable for (dichotomous) qualitative tests and quantitative tests categorized into two or more categories. - In general, the acceptance criteria for accuracy is >90%.
Kappa	A statistic method to measure agreement between two tests, that take into account that the agreement occurs simply by chance.	<table><tr><td></td><td></td><td colspan="2">**Reference test**</td><td></td></tr><tr><td></td><td></td><td>+</td><td>–</td><td></td></tr><tr><td>**New test**</td><td>+</td><td>TP</td><td>b</td><td>g1</td></tr><tr><td></td><td>–</td><td>FN</td><td>d</td><td>g2</td></tr><tr><td></td><td></td><td>f1</td><td>f2</td><td>N</td></tr></table> *Po* = observed agreement $(a + d)/N$, *Pe* = agreement expected by chance $[(g1 \times f1) + (g2 \times f2)]/N^2$.	- Alternative measure of accuracy in validation and verification. - Can be used for qualitative tests and semi-quantitative tests with two or more possible categories. - Kappa needs an interpretation table <0: poor 0.01–0.20: slight, 0.21–0.40: fair, 0.41–0.60: moderate, 0.61–0.80: substantial, and 0.81–1.00: almost perfect. - Affected by the skewed distribution of positive and negative reference tests (lower kappa in skewed distribution) [30].
Precision (reproducibility)	Closeness of agreement between results of replicate measurements (of the new test)	- Intra-test precision: selected samples are tested several times at once (single run). - Inter-test precision: selected samples are tested several times on multiple days. - Calculation For quantitative test (precision): coefficient of variation = (standard deviation of measurements/mean) × 100. For qualitative test (reproducibility): (Number of repeated results in agreement/total number of results) × 100.	- Another important parameter for acceptance next to accuracy in validation and verification [2]. - In general, reproducibility of >90% is needed for acceptance. - For quantitative (precision) and qualitative (reproducibility) tests. - For quantitative test, precision should be performed at at least two levels (high titer and low titer). - For qualitative test, reproducibility should include at least one positive and one negative sample. - Using samples samples around cut-off may reduce precision calculation of semi-quantitative tests. - It can be done by performing the new test in triplicate in 1 day and repeat this triplication for 5 days.
Clinical sensitivity	The ability of a test to detect the presence of the disease of interest correctly.	- Create a 2 × 2 contingency table. <table><tr><td></td><td></td><td colspan="2">**Clinical diagnosis**</td></tr><tr><td></td><td></td><td>+</td><td>-</td></tr><tr><td>**New test**</td><td>+</td><td>TP</td><td>FP</td></tr><tr><td></td><td>-</td><td>FN</td><td>TN</td></tr></table> [True positive (TP)/(TP + false negative (FN)].	- Used in validation and should be interpreted together with the test's role in the diagnostic cascade. - Clinical data are often not available as reference standards. - When the sensitivity is calculated using previous tests as a reference standard, perhaps the correct term should be diagnostic sensitivity. Yet, this sensitivity is not the sensitivity of the test that differentiates disease from no disease. - It is used for dichotomous qualitative tests and semi-quantitative tests (quantitative tests categorized into two possible variables). - A test with high clinical sensitivity is needed for a screening test.
Clinical specificity	That ability of a test to correctly the absence of the disease of interest.	- Create a 2 × 2 contingency table. - [True negative (TN)/TN + false positive (FP)]	- In our opinion, it should be used in validation and should be interpreted together with the test's role in the diagnostic cascade. - See also the comments above.
Analytical sensitivity	The ability of a test to detect the lowest concentration of target (e.g. antibody, antigen, and microorganism).	- By testing multiple specimens across several dilutions. - Two measures: limit of detection (LOD) and limit of quantification (LOQ). - LOD: the lowest concentration of target that still can be detected in the test. - LOQ: the lowest concentration of target that still can be reliably measured by the test.	- It should be mentioned in validation [2]. - It should not be confused with clinical or diagnostic sensitivity. - Applicable for quantitative tests.
Analytical specificity	The ability of a test to detect only the intended target.	- By adding interference substance (e.g. other targets) to a certain number of samples that are tested. - True negatives/(true negatives + false positives).	- It should be mentioned in validation [2]. - It should not to be confused with clinical or diagnostic specificity.
Linearity	The determination of whether measured results are directly proportional to the concentration of the target.	- By testing multiple specimens across several dilutions. - Measured as deviation from an ideal straight line.	- It should be mentioned in the validation of quantitative tests.

Table 2 (*continued*)

	Definition	How to measure and calculate	Remarks and explanation
Essential agreement	Percentage of isolates with the MIC as measured by the test within ±1 doubling dilution of the corresponding MIC as measured by reference standard [18,19].		- This is the parameter for accuracy in AST verification. It should be used when possible (i.e. not for disk diffusion because it does not report MIC). This term is used in both CLSI M52 [18] and ISO 20776-2:2021 [19] documents. - In our opinion, this is the most important acceptance criterion for the AST verification procedure. - ≥90% is the essential agreement criteria for acceptance, meaning in 30 isolates, only three results outside the essential agreement can be accepted [18]. - It should be calculated per antibiotic. - The reference AST method should be the same as the test under investigation (i.e. gradient test vs. gradient test). Otherwise, there is a risk that the EA and CA are lower, not because of the performance of the AST, but because of methodological comparison.
Categorical agreement	The proportion of isolates that are classified in the same susceptibility category by the test as the susceptibility category determined by the reference standard [18].		- Another parameter for accuracy in AST verification. - This term is used in CLSI M52 [18] but not in ISO 20776-2:2021 [19] documents. The argument for ISO 20776-2:2021 not to use this term is that the verification is aimed at measures of AST performance only and does not result in interpretation using breakpoints. - It should be calculated per antibiotic. - ≥90% is the categorical agreement criteria for acceptance, meaning in 30 isolates, only three results outside the categorical agreement can be accepted [18]. - Can be performed only for AST that reports MIC and breakpoint is available. - The same breakpoint methods (i.e. EUCAST or CLSI) and version should be used for the new and reference AST. - It should take into account that using isolates with the MIC around breakpoints may reduce CA.
Very major discrepancy or very major error	The proportion of the isolates deemed as susceptible by the AST under investigation while the reference standard reports resistance [18].		- Another parameter for accuracy in AST verification. - This term is used in both CLSI M52 but not in ISO 20776-2:2021 documents. - It should be calculated per antibiotic. - CLSI uses acceptable very major discrepancy/error of <3%,with number of isolates resistant by the reference method as denominator. It means in 30 isolates <1 very major discrepancy can be accepted [18]. Using total number of tested isolates as denominator allows 1 very major error [$(1/30) \times 100 = 3.3\%$]. - Very major discrepancy as a term is used when the reference standard is other AST that is previously used in the lab and very major error should be used when the reference standard is BMD.
Major discrepancy or major error	The proportion of the isolates deemed as resistant by the AST under investigation, among isolates deemed as susceptible by the reference method [18].		- Another parameter for accuracy in AST verification. - This term is used in both CLSI M52 but not in ISO 20776-2:2021 documents. - It should be calculated per antibiotic. - CLSI uses acceptable major discrepancy/error of <3% with number of isolates susceptible by the reference method as denominator. [18]. However, using 30 isolates in which the half is susceptible (n=15), the allowed major error is <1, Using total number of tested isolates as denominator, allows only 1 major error [$(1/30) \times 100 = 3.3\%$]. - In our opinion, this is really strict. When ME is considered as acceptance criteria, higher number should be acceptable [i.e. 3 isolates with major discrepancy/error),. - Major discrepancy as the term is used when the reference standard is another AST that is previously used in the lab and major error should be used when the reference standard is BMD.
Reproducibility of antimicrobial susceptibility test	Three isolates (reference microorganisms such as ATCC or clinical isolates) were tested three times a day for 5 days [18].		

AST, antimicrobial susceptibility test; ATCC, American Type Culture Collection; BMD, broth microdilution; CA, Categorical agreement; CLSI, Clinical & Laboratory Standards Institute; EUCAST, European Committee on Antimicrobial Susceptibility Testing; ISO, International Organization for Standardization; MIC, minimum inhibitory concentration.

Table 3
Number of samples needed for validation and verification purposes

Purpose	Suggested numbers of minimum isolates needed from various sources, together with references	Arguments and comments
Validation	≥50 positive specimens and ≥100 negative specimens [6]. Own calculation using information from [31,32] 120 (when standardized mean difference of 0.5 is used). 50 (when a standardized mean difference of 0.8 is used).	No comments or calculations were given as to why the authors from reference [6] suggest this sample size for validation purposes. Using the Altman normogram takes into account power (that represents the sensitivity of a test) and standardized difference (minimum difference divided by standard deviation for a quantitative test), the number of samples can be calculated. A power of 80% is often arbitrarily chosen as power in diagnostic studies. A standardized difference of 0.5 is often considered as medium and 0.8 as large [33]. When a clinical reference standard is used and the new test is run prospectively in parallel with the reference standard using clinical samples, the distribution of positive and negative samples will depend on the prevalence of the disease in the specific population. If stored or spiked samples are used instead, two-thirds of the samples should be positive.
Verification	≥20 samples (generally divided equally among positive and negative samples) [6]. ≥20 positive and 50 negative samples [34].	Flexibility in the distribution of positives and negatives is important to account for such factors as rare analytes and the institution's patient population. Caution should be exercised when selecting only 10 positive and 10 negative samples. In such cases, the occurrence of even a single false positive can markedly diminish the sensitivity value.
Calculation of reproducibility	At least two positive and two negative samples [6]. For the quantitative test: at least two samples at two levels (low and high concentration).	
Verification of antimicrobial susceptibility test	≥30 [18].	Should be representative of isolates in the laboratory and diverse enough in terms of resistance mechanisms. Effort should be made to include ≥50% of the isolates that are non-wild type, and include major resistance mechanisms encountered in the laboratory.
Reproducibility of antimicrobial susceptibility test	Five isolates were tested three times [19].	Quality control isolates (ATCC) or clinical isolates can be used.

AST: antimicrobial susceptibility test; ATCC: American Type Culture Collection.

Number of samples	95% confidence interval
50	66.3% to 90.0%
100	70.8% to 87.3%.

Fig. 2. Confidence intervals vary with sample size. The table shows the confidence interval when validating a test with 80% sensitivity in 50 samples and 100 samples.

cut-off, or microorganisms that are rarely encountered in the laboratory, may be included because it will strengthen the quality of the verification.

Collecting enough samples and isolates may require assistance from other or reference laboratories, such as collecting positive urine samples with *Histoplasma* in a non-endemic setting. However, this process can still be lengthy and it may be necessary to reconsider whether a validation procedure should be performed.

Using documented test results as a comparison

Sometimes, verification is performed by comparing the new test with the documented results of the reference standard. This approach is efficient as only one test should be performed at a time. However, differences in test conditions and handling can cause discrepant results.

Making a decision

The results of the new test are compared with those of the reference standard using a contingency table, allowing the calculation of parameters to be compared with the pre-defined criteria (Table 2). Accuracy is the most important acceptance criterion, with >90% generally required [6] to accept the test to be used. Yet, it is flexible for various reasons. Accuracy alone provides less information without reporting sensitivity and specificity separately (Fig. 3). A more sensitive new test may appear less specific if the reference standard labels positive results as false positive. Sometimes, a high accuracy should be accompanied by high sensitivity (>95%) when microorganism detection is important. For example, a chromogenic agar detecting methicillin-resistant *Staphylococcus aureus* (Class C IVDR) for screening purposes, needs to have high sensitivity next to high accuracy. Often, Cohen's kappa is used as an accuracy parameter. It takes into account that the agreement between two tests occurs by chance. Yet, we advise against using kappa in verification due to its difficult interpretation and arbitrary cut-offs. In addition, high agreement can be accompanied by low kappa when the distribution of the number of positive and negative reference tests is skewed [13].

Accuracy is influenced by the inclusion of 'difficult samples' that are used to investigate the robustness of the test. Simply by nature, samples with results near the cut-off value may yield positive or negative results upon repetition. Therefore, a plan should be made before verification, to exclude these samples from accuracy calculations or to lower acceptance criteria. This decision needs to be transparent and documented in the verification report (Table 4).

Example calculation of verification study of a test with high accuracy but rather low sensitivity. When a new test is a test that is planned to be used as a screening test (e.g. MRSA chromogenic agar), despite having an acceptable accuracy (≥90%), the test should not be accepted for use due to its low sensitivity.

		Reference test	
New test		**Positive (n)**	**Negative (n)**
	Positive (n)	15	2
	Negative (n)	5	48

Accuracy ((15+48) / (15+2+5+48)) x 100% = (63/70) x 100% = 90.0%

Sensitivity (15/20) x 100% = 75.0%

Example calculation of verification study of a test with high, acceptable accuracy (≥90%). Yet, when the new test needs to be highly specific (i.e. serum galactomannan antigen test for the diagnosis of invasive fungal infection) due to potential significant consequence (long term antifungal therapy), it should be discussed whether this test with this verification performance should be accepted.

		Reference test	
New test		**Positive (n)**	**Negative (n)**
	Positive (n)	20	7
	Negative (n)	0	43

Accuracy ((20+43) / (20+7+0+43)) x 100% = (63/70) x 100% = 90.0%

Specificity (43/50) x 100% = 86.0%

Fig. 3. Examples of calculations where discrepancies can occur between accuracy and sensitivity or specificity. In these cases, despite high and acceptable accuracy, further judgement is needed to determine whether the test should be accepted to be used in the laboratory.

Table 4
Steps in performing a method validation or verification of a diagnostic microbiology test. All of these steps should be documented and written in a report

1. Determine why a new test is needed (Better performance? Cheaper? Less laborious? Higher volume? Safety?)
2. If > 1 tests are available, look at other factors such as cost, labour intensive, speed, and safety.
3. What is the test intended to do? (Screening? Diagnostic? Confirmation?)
4. Look at the performance of the test according to the package insert or other publications.
5. When there are several CE-labelled tests are available, choose which test(s) will be validated/verified.
6. Write the protocol before starting the validation or verification process
   - Choose the reference standard.
   - Decide on criteria to accept the test (this should be done before the validation or verification process).
   - Determine the number of samples to be used for the validation or verification process.
   - How the validation or verification should be done? Options are among others: (i) prospective in parallel (new and reference test in patients samples), (ii) prospective in parallel in stored samples, and (iii) prospective only for new test, and old, readily documented results from the reference standard is used as comparison.
7. Perform the validation or verification process.
8. Write the results.
9. Calculate the accuracy and reproducibility.
   - Solve the discrepancy.
10. Make a decision on whether to accept or reject the test to be used in the lab based on pre-defined criteria. Another option is to postpone the decision because more data are needed.
11. When more data are needed, discuss the results of validation or verification with stakeholders.

Furthermore, we recommend using >90% reproducibility as the acceptance criterion, as high trust in the test is essential. Other parameters than accuracy and reproducibility mentioned in Table 2 are usually not used to determine acceptance of a test but necessary to have a comprehensive description of the new test.

Sometimes, a decision to implement a test cannot be made after verification. For example, a new *Aspergillus* antigen test might be easier to perform than a previous test and has acceptable accuracy but lower specificity. False positive results of new test can lead to unnecessary and prolonged antifungal treatment. Conversely, the older test might miss true positives, resulting in missed diagnoses of invasive pulmonary aspergillosis, a disease with significant mortality. In such a case, clinical microbiologists and clinicians should discuss the verification results and take clinical needs to implement the test into consideration, and decide together whether to implement the new test.

Solving discrepancy

Discrepancies between the new test and reference standard results should be examined. These are often resolved by a third method (another test or clinical criteria) or by repeating the test on discrepant samples. When this is done, it is important to recognize potential bias, as tests are performed only on discordant samples. The third test may

also be inferior to the reference test. In addition, if the diagnosis is based on multiple criteria, any of the tests in the comparison may be part of those criteria, leading to biased results. For example, a verification study comparing a new *Aspergillus* antigen test with a reference standard (another antigen test) may try to solve discrepancies by using the European Organization for Research and Treatment of Cancer and Mycoses Study Group (EORTC-MSG) criteria of probable invasive aspergillosis [14]. However, as the microbiology criterion is part of the set of diagnostic criteria, this creates circular reasoning. Another option to resolve discrepancies is to use a composite reference standard [15,16]. This can be done in the case of verifying more than two tests simultaneously, in which the most frequent results is considered as the true value.

Antimicrobial susceptibility tests

Automated ASTs have a CE label and are classified as Class B tests [17], but CE label information for disks and gradient tests is often unavailable. Agar plates used for manual ASTs are interpreted by some manufacturers as Class A tests [17]. The Clinical & Laboratory Standards Institute (CLSI) M52 and ISO 20776-2:2021 on verification of ASTs, focus on automated AST [18] and MIC-producing ASTs [19], respectively. The latter specifically addresses verification using reference broth microdilution (BMD). We believe BMD as a reference method should be reserved for verifying commercial BMD ASTs because of inherent technical differences between methods as often observed when comparing different AST methods [20,21]. Such discrepancy may be specific to a certain AST method in combination with certain microorganisms [22]. European Committee on Antimicrobial Susceptibility Testing (EUCAST) itself uses BMD to correlate disk diffusion breakpoints. In addition, not all laboratories have the capacity to prepare BMD according to this ISO standard.

It is argued that the laboratories can perform disk diffusion and gradient diffusion, so these manual ASTs do not need verification [23]. Yet, ASTs with CE labels still should be verified because of potential influences from transport, handling, and using different Mueller–Hinton agar than what manufacturers used in their validation [24,25]. Validation, which is more extensive and involves more samples than verification, is rare in laboratories since they seldom develop their own ASTs.

The CLSI recommends using a minimum of 30 isolates [18] and based its calculation on probability to meet the acceptance criterion for essential agreement. The isolates should be diverse and reflect the prevalence of microorganisms isolated in the laboratory. For instance, when verifying a new gram-negative card for an automatic AST, the following isolates may be selected: *Escherichia coli* ($n = 10$), *Klebsiella pneumoniae* ($n = 5$), *Enterobacter cloacae* ($n = 3$), *Proteus mirabilis* ($n = 3$), *Pseudomonas aeruginosa* ($n = 4$), and other Enterobacterales (each one). With only 30 isolates, it is impossible to test a new AST's accuracy for all resistance mechanisms, but we suggest including at least extended spectrum beta-lactamases (ESBLs), ampC, various carbapenemases producing Enterobacterales (CPEs) (OXA-48, KPC, and NDM), gentamicin- and fluoroquinolone-resistant isolates in at least half of the selected isolates. In Belgium, a panel of 14 gram-positive and 14 gram-negative isolates covering a wide spectrum of resistance mechanisms has been developed [26]. The CLSI M52 document also mentions using 10 samples for limited verification when minor changes occur in the AST system (e.g. software change) [18]. However, we believe this is unnecessary, since 10 isolates is not representative for the isolates isolated in the laboratory and possible error will remain undetected despite this limited verification. Performing limited verification for every minor changes is also impractical and laborious for the laboratory to perform. Moreover, when one or two errors occur simply by chance in a small sample size, this can lead to unnecessary further investigation and delay updates.

It is important to test new and reference AST methods in parallel using the same bacteria inoculum [16] and not use previously determined MIC results as reference standards. This is because test conditions between the new and reference standards differ, whereas variation in conditions should minimized during verification [27]. Unlike CLSI M52, ISO 20776-2:2021 [19] does not use the term categorical agreement (CA) (Table 2) as accuracy measurement. It argues that verification should measure test performance and not the result of interpretative reading using breakpoints (i.e. categorizing MICs into susceptible or resistant). Yet, because interpretative reading is widely used in routine, we favour to report CA, (very) major error (VME), next to an essential agreement (EA) as in CLSI M52 [19]. These measures will indicate the direction of potential biases of the new test. VME will have more serious clinical consequences than major error (ME). However, in our opinion, the most important acceptance criteria should be EA. In a verification procedure using 30 isolates as advised by CLSI and using CLSI acceptance criteria, the maximum allowed non-CA would be three and only one VME or one ME is allowed. Including isolates with MIC around breakpoints may also lower the CA. The misclassification may occur simply because of the measurement nature (the same isolate is sometimes categorized as susceptible, sometimes as resistant). Errors should be investigated and resolved using a third AST method (our preference), a molecular test, or by repeating the new test.

Concluding remarks

The introduction of IVDR may increase verification procedures as laboratories transition from LDTs to commercial tests, and some may choose to validate their LDTs. In this paper, we have summarized, interpreted, and given arguments on how to perform these procedures.

Author contributions

Study concept and design: all authors. Data acquisition and analysis: EY and ML. All authors contributed to the interpretation of data. Drafting of the manuscript: EY. All authors critically revised the manuscript for intellectual content and approved the final draft for submission.

EY and MS are clinical microbiologists with over 8 years of experience in ISO-accredited laboratories in the Netherlands, with EY working in an academic setting and MS in a non-academic setting. They have conducted numerous validation and verification procedures, many of which have been published in peer-reviewed journals. EY also has previous experience working in academic and non-academic clinical microbiology laboratories in Belgium. Both are expert members of the Dutch Foundation for Quality Assessment in Medical Laboratory Diagnostics. ML's research focuses on the methodology of medical test evaluations, including diagnostic test accuracy, the efficacy of tests, and policy advice on screening tests.

Transparency declaration

EY is a paid external expert for Dekra, a notified and certification body for medical devices. No funding was received for this study.

Acknowledgements

We thank Maarten Heuvelmans, MD (clinical microbiologist, Medical Microbiology and Immunology Gelderland (MMIG), and member of quality commission of Dutch Medical Microbiology Association (NVMM)), Martijn den Reijer, MD, PhD (clinical microbiologist, Star-Shl diagnostic laboratory, Rotterdam, The Netherlands), Melissa Depypere, MD, PhD (clinical biologist, Leuven University Hospital, Belgium), and Anna Tisler, MD, PhD (clinical microbiologist, Institute of Family Medicine and Public Health, University of Tartu, Estonia) for their invaluable comments on the manuscript.

References

[1] European Commission. 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. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32017R0746.

[2] ISO. ISO 15189:2022(en). Medical laboratories—requirements for quality and competence. https://www.iso.org/obp/ui/en/#iso:std:iso:15189:ed-4:v1:en. [Accessed 26 June 2023].

[3] Gilbert R, Logan S, Moyer VA, Elliott EJ. Evidence-based case review: assessing diagnostic and screening tests: Part 1. Concepts. West J Med 2001;174:405. https://doi.org/10.1136/ewjm.174.6.405.

[4] Tibbetts RJ. Verification and validation of tests used in the clinical microbiology laboratory. Clin Microbiol Newsl 2015;37:153–60. https://doi.org/10.1016/j.clinmicnews.2015.09.004.

[5] Tan TY, Jiang B, Ng LSY. Faster and economical screening for vancomycin-resistant enterococci by sequential use of chromogenic agar and real-time polymerase chain reaction. J Microbiol Immunol Infect 2017;50:448–53. https://doi.org/10.1016/j.jmii.2015.08.003.

[6] Clark RB, Lewinski MA, Loeffelholz MJ, Tibbetts RJ. In: Sharp S, editor. Cumitech 31A, verification and validation of procedures in the Clinical Microbiology laboratory, 24. Washington, DC: ASM Press; 2009.

[7] Lord SJ, St John A, Bossuyt PMM, Sandberg S, Monaghan PJ, O'Kane M, et al. Setting clinical performance specifications to develop and evaluate biomarkers for clinical use. Ann Clin Biochem 2019;56:527–35. https://doi.org/10.1177/0004563219842265.

[8] Doust JA, L Bell KJ, G Leeflang MM, Dinnes J, Lord SJ, Mallett S, et al. Guidance for the design and reporting of studies evaluating the clinical performance of tests for present or past SARS-CoV-2 infection. BMJ 2021;29:n568. https://doi.org/10.1136/bmj.n568.

[9] Peter Donnelly J, Chen SC, Kauffman CA, Steinbach WJ, Baddley JW, Verweij PE, et al. Revision and update of the consensus definitions of invasive fungal disease from the European organization for research and treatment of cancer and the mycoses study group education and research consortium. Clin Infect Dis 2020;71:1367–76. https://doi.org/10.1093/cid/ciz1008.

[10] Walsh T. Fuzzy gold standards: approaches to handling an imperfect reference standard. J Dent 2018;74:S47–9. https://doi.org/10.1016/j.jdent.2018.04.022.

[11] Leeflang MMG, Allerberger F. Sample size calculations for diagnostic studies. Clin Microbiol Infect 2019;25:777–8. https://doi.org/10.1016/j.cmi.2019.04.011.

[12] Chu H, Cole SR. Sample size calculation using exact methods in diagnostic test studies. J Clin Epidemiol 2007;60:1201–2. https://doi.org/10.1016/j.jclinepi.2006.09.015.

[13] de Vet HCW, Mokkink LB, Terwee CB, Hoekstra OS, Knol DL. Clinicians are right not to like Cohen's κ. BMJ 2013;346:f2125. https://doi.org/10.1136/bmj.f2125.

[14] Calero AL, Alonso R, Gadea I, Vega MDM, García MM, Muñoz P, et al. Comparison of the performance of two galactomannan detection tests: platelia *Aspergillus* Ag and *Aspergillus* galactomannan Ag VirClia Monotest. Microbiol Spectr 2022;10:e02626. https://doi.org/10.1128/spectrum.02626-21. 21.

[15] Hadgu A, Miller W, Pepe MS, Alonzo TA. Using a combination of reference tests to assess the accuracy of a diagnostic test (multiple letters). Stat Med 2001;20:656–8. https://doi.org/10.1002/(sici)1097-0258(19991130)18:22<2987::aid-sim205>3.0.co;2-b.

[16] Yusuf E, Van Westreenen M, Goessens W, Croughs P. The accuracy of four commercial broth microdilution tests in the determination of the minimum inhibitory concentration of colistin. Ann Clin Microbiol Antimicrob 2020;19:42. https://doi.org/10.1186/s12941-020-00383-x.

[17] bioMerieux. References CE marked under IVDR. https://resourcecenter.biomerieux.com/settings/ivdr/ivdr.pdf. [Accessed 28 June 2024].

[18] CLSI. Verification of commercial microbial identification and susceptibility test systems, M52 guideline. Wayne, PA: Clinical and Laboratory Standards Institute; 2016.

[19] The International Organization for Standardization. ISO 20776-2:2021 Clinical laboratory testing and *in vitro* diagnostic test systems—susceptibility testing of infectious agents and evaluation of performance of antimicrobial susceptibility test devices. ISO Int Organ Stand 2021;20776. ISO, Geneva, Switzerland, https://www.iso.org/obp/ui/en/#iso:std:iso:20776:-2:ed-2:v1:en.

[20] Liao CH, Kung HC, Hsu GJ, Lu PL, Liu YC, Chen CM, et al. Activities of tigecycline against clinical isolates of *Acinetobacter baumannii* in Taiwan: broth microdilution method vs. disk diffusion method. Int J Infect Dis 2008;12:e406. https://doi.org/10.1007/s10096-020-04121-1.

[21] Luber P, Bartelt E, Genschow E, Wagner J, Hahn H. Comparison of broth microdilution, E test, and agar dilution methods for antibiotic susceptibility testing of *Campylobacter jejuni* and *Campylobacter coli*. J Clin Microbiol 2003;41:1062. https://doi.org/10.1128/JCM.41.3.1062-1068.2003.

[22] EUCAST. Area of Technical Uncertainty (ATU) in antimicrobial susceptibility testing. EUCAST. https://www.eucast.org/eucast_news/news_singleview?tx_ttnews%5Btt_news%5D=297&cHash=d8147b987630472e894225f3780a794e. [Accessed 28 June 2024].

[23] Kirby JE, Brennan-Krohn T, Smith KP. Bringing antimicrobial susceptibility testing for new drugs into the clinical laboratory: removing obstacles in our fight against multidrug-resistant pathogens. J Clin Microbiol 2019;57:e01270. https://doi.org/10.1093/cid/ciz089. 19.

[24] Patel JB, Thomson RB, Alby K, Babady E, Culbreath K, Galas MF, et al. Expert opinion on verification of antimicrobial susceptibility tests. J Clin Microbiol 2020;58:945–65. https://doi.org/10.1128/JCM.00945-20.

[25] Humphries RM, Simner PJ. Verification is an integral part of antimicrobial susceptibility test quality assurance. J Clin Microbiol 2020;58:1986–2005. https://doi.org/10.1128/jcm.01986-19.

[26] Desmet S, Verhaegen J, Glupzcynski Y, Van Eldere J, Melin P, Goossens H, et al. Development of a national EUCAST challenge panel for antimicrobial susceptibility testing. Clin Microbiol Infect 2016;22:704–10. https://doi.org/10.1016/j.cmi.2016.05.011.

[27] Hindler JA, Humphries RM. Colistin MIC variability by method for contemporary clinical isolates of multidrug-resistant gram-negative bacilli. J Clin Microbiol 2013;51:1678–84. https://doi.org/10.1128/jcm.03385-12.

[28] Public Health England. UK standards for microbiology investigations: Q1. Evaluation, validation and verifications of diagnostic tests UK standards for microbiology investigations. 2017. p. 45.

[29] Ilstrup DM. Statistical methods in microbiology. Clin Microbiol Rev 1990;3:219. https://doi.org/10.1128/cmr.3.3.219.

[30] Dettori JR, Norvell DC. EBSJ special section: science-in-Spine. Kappa and beyond: is there agreement?. https://us.sagepub.com/en-us/nam/open-access-at-sage. [Accessed 4 May 2023].

[31] Altman DG. Medicine and mathematics statistics and ethics in medical research III How large a sample? Br Med J 1980;281:1336–8. https://doi.org/10.1136/bmj.281.6251.1336.

[32] Columb MO, Fficm F, Atkinson MS. Statistical analysis: sample size and power estimations. https://access.oxfordjournals.org. [Accessed 8 February 2024].

[33] Luo Yan, Funada S, Noma H, Furukawa TA. How is standardized mean difference computed, reported and interpreted in randomized controlled trials: protocol for a meta-epidemiological study. https://doi.org/10.17605/OSF.IO/G9RDH.

[34] Elder BL. Verification and validation of procedures in the clinical microbiology laboratory. Clin Microbiol Newsl 1997;19:153–6. https://doi.org/10.1016/S0196-4399(97)83919-0.