Clinica Chimica Acta 497 (2019) 197–203

Contents lists available at ScienceDirect

Clinica Chimica Acta

journal homepage: www.elsevier.com/locate/cca

Discussion

Evaluation of analytical performance of qualitative and semi-quantitative assays in the clinical laboratory

Joachim K.W. Pum

Bioscientia MVZ Jena, Orlaweg 2, 07743 Jena, Germany

ARTICLE INFO

Keywords:
Qualitative assays
Semi-quantitative assays
Bayesian statistics
Cohen's kappa
Weighted kappa
Validation
Verification
Coefficient of unalikeability

ABSTRACT

As is true for quantitative assays, qualitative and semi-quantitative assays, producing strict binary or ordinal results, must undergo a verification process prior to their implementation for routine clinical laboratory testing. Standard method validation parameters used for quantitative assays, however, do not apply here. Rather, contingency tables, Bayesian statistics and statistical hypothesis testing for inter-rater agreement must be used. This article provides an overview of simple, practical tools, which can be used to verify the analytical performance of such assays. Topics discussed include the verification of precision and accuracy with a single experiment approach and performing method comparison experiments for assays with binary or ordinal results. Acceptance criteria are recommended for each test to provide a standardized framework for performance assessment. The approach is appropriate for all CE/IVD-marked and CLIA-waived assays and will ensure compliance with CAP, ISO 17025 and ISO 15189 regulations.

1. Introduction

Qualitative assays are used for a variety of applications in the clinical laboratory, ranging from screening to diagnosis and disease monitoring. The technologies employed are diverse and can include simple methods, such as visual readings, up to highly complex methods, such as automated nucleic acid sequencing or micro-arrays [1]. Results of qualitative assays are usually nominal, meaning that they can be either positive or negative (only a binary condition is known). A result is reported as “positive” when it exceeds a predefined cutoff value and as “negative” when the cutoff-value is not reached. Depending on the test design and the clinical objectives, different cutoffs can be chosen, to minimize either the number of false-positive or false-negative results. Correspondingly, the diagnostic sensitivity and specificity of such assays can vary considerably.

Semi-quantitative assays are not semi-quantitative in the sense of having a large uncertainty, but rather ordinal, meaning that more than two alternative values are possible [2]. A typical example for such test procedures are urine test strips, ABO blood groups, pregnancy tests, screening tests for drugs of abuse and qualitative molecular tests [3,4]. In all of these test systems measurement results are categorized in multiple arbitrary steps with result categories without measurement units [5]. Like quantitative assays, qualitative and semi-quantitative assays must undergo a verification process prior to their introduction in the clinical laboratory. As pointed out in the CLSI document EP12-A2, owing to the large diversity of qualitative assays, universal evaluation guidelines are not feasible. However, common features do exist and therefore precision studies and method comparison studies should always make up part of the verification process [1]. The ISO standards DIN EN ISO 15189 and DIN EN ISO/IEC 17025 specify that methods, which have been validated by the manufacturer and are implemented without modification, must have precision and bias verified. Furthermore, a method comparison study must be carried out when replacing an existing method with a new one. Methods, which have not been validated by the manufacturer or standard methods, which are used for applications not intended by the manufacturer must be validated to such an extent, as to satisfy the requirements of the given application [6,7]. Similarly, the “All Common Checklist” of the College of American Pathologists (CAP) Accreditation Program states that, “the laboratory must verify or establish the method performance specifications that are applicable and clinically relevant” [8]. For qualitative and semi-quantitative FDA cleared tests, this typically includes verifying accuracy, precision, and carrying out a method comparison study. For non-FDA cleared tests, diagnostic sensitivity and diagnostic specificity must also be established [9,10].

The goal of this article is to provide an overview of tools available for the evaluation of qualitative and semi-quantitative assays. It is assumed that the vast majority of assays employed in the routine laboratory have been previously validated by the manufacturer and are implemented without modifications. Formal validation of diagnostic

E-mail address: joachim.pum@bioscientia.de.

https://doi.org/10.1016/j.cca.2019.07.018
Received 18 June 2019; Received in revised form 15 July 2019; Accepted 16 July 2019
Available online 21 July 2019
0009-8981/

accuracy by the user is therefore not required. Nevertheless, precision and analytical accuracy must be verified and a method comparison experiment must be carried out if CAP, ISO 15189 or ISO 17025 requirements are to be satisfied. While tools such as *likelihood ratios*, *odds ratios* and *receiver operating characteristic (ROC)* are extremely valuable in characterizing diagnostic tests, they refer to *diagnostic accuracy* and will therefore not be discussed further here. Information on these is readily available in standard textbooks on medical statistics and beyond the scope of this article. Rather, this review will focus on practical tools, which can be used to characterize the *analytical performance* of such assays, e.g. calculating precision and accuracy for qualitative and semi-quantitative assays and performing a method comparison study. Finally, acceptance criteria are recommended for each test, in an attempt to provide the analyst with a standardized framework, against which to assess analytical assay performance.

2. Verification of method performance

2.1. Precision and accuracy

Precision, or random error, is defined by ISO 3534-1 as "a measure of the closeness (degree of scatter) between independent test results obtained under stipulated conditions" [11] and is usually expressed numerically by measures of imprecision, such as standard deviation or relative standard deviation (coefficient of variance) of repeat measurements. Many assays, such as immunological screening assays for drugs-of-abuse or serology assays for viral pathogens, commonly produce numerical values, which are converted to qualitative results, based on one or more cutoffs. In such cases, precision can be estimated in an analogous fashion to quantitative assays, using measurement signals instead of measurement results. As it is common for precision to vary with analyte concentrations it should be determined at two or more concentration levels, if the analyte concentration is expected to vary by > 50% of an average value [12]. Typically, certified reference materials, quality control samples or aliquots of independently prepared laboratory sample pools are used for this purpose.

When dealing with assays that produce strict binary or ordinal results, a different approach must be taken, as variability in categorical data is different from variability in quantitative data. Rather than measuring by how *much* the observations differ from the mean, the focus is shifted to determining how *often* the observations differ. A number of concepts are available for this purpose, including accordance and concordance [13–15], an iterative procedure described by Mandel [16] and a method to determine precision by interlaboratory experiments described by Wilrich [17]. While each of these procedures has its own merits, the main disadvantages lie in the fact that they are either not equivalent to repeatability / reproducibility, contain complex calculations or require measurements from multiple sites. Moreover, these concepts are applicable only to binomial data. Kader and Perry introduced a simple parameter, called the *"Coefficient of Unalikeability (CU)"*, which provides a quantitative method for the measurement of variability for categorical variables, is easy to calculate and can be applied to binomial and ordinal data [18,19]. It is defined as the ratio of the total number of differences in a distribution to the maximum number of possible differences in the same distribution and can be related to the widely used Bernoulli variable in terms of the intra-data variability described by Gordon [18,20]. Thus, if all the responses are in the same category, there is no variability and the CU is 0. In contrast, when the cases are distributed uniformly across the categories, variability is at its maximum and CU is 1.

Accuracy is the quantitative expression of trueness, which is defined as the "closeness of agreement between the average of an infinite number of replicate measured quantity values and a reference quantity value" [21]. It is commonly expressed in terms of bias, which may be constant or proportional and is either positive or negative, as compared to the true value [22]. It can be calculated as the mean difference between results obtained with a candidate method and an existing (reference) method or as the mean difference between results obtained by replicate measurements of a reference material and the assigned target value of the reference material. Again, measurement signals can be used to carry out calculations analogously to quantitative methods, providing that numerical results precede the qualitative output of the method. When dealing with methods, that provide strict binary or ordinal results, a different approach is once again called for to estimate appropriate performance parameters.

While there is nothing to be said against estimating precision and accuracy in separate experiments, a single experiment approach is usually preferred, as it offers significant advantages in terms of efficiency and computational rigor.

2.1.1. Qualitative assays

When verifying precision and accuracy of qualitative assays with a single experiment, a minimum of 20 samples with known concentration (e.g. quality control samples, certified reference materials or patient samples) should be measured over a period of at least five days [23,24]. Typically, one positive, one weak-positive and two negative control samples are measured per day and results summarized in a 2 × 2 contingency table (Table 1). The weak-positive samples are included to ensure that the system is adequately challenged and should have a concentration of approximately 20% above the cutoff. This approach represents a reasonable compromise between costs associated with time and materials and a valid assessment of the procedure's performance.

Symmetry testing: The first step is to examine the data for a systematic difference between the results obtained and the target values for the samples. This is assessed with McNemar's test, which is a modification of the ordinary chi-square test for paired data. It uses values from the two discordant cells b and c (see Table 1) to test the equivalence of the two proportions (i.e., marginal homogeneity). A limitation of the McNemar test is that it was developed for large samples. For small sample sizes ($n \leq 25$) a correction formula, such as the Yates correction formula, should be used [25]. Furthermore, hypothesis testing is based on a chi-squared distribution and assumes that the sum of the discordant pairs ($b + c$) is ≥10. If this is not true, an exact binomial test should be used [26]. The test statistic is calculated as:

$$\frac{(b-c)^2}{(b+c)} \tag{1}$$

and with the Yates correction:

$$\frac{(|b-c|-1)^2}{(b+c)} \tag{2}$$

A statistically significant value (e.g. $p < 0.05$) indicates a substantial number of false-positive or false-negative results and will need to be investigated further.

Precision: The percent coefficient of unalikeability (*CU*%) is calculated separately for the positive and negative control sample:

$$CU\% = 2p_1p_2 \times 100 \tag{3}$$

Table 1
Comparison data between a reference method (or target values of a control sample or certified reference material) and a comparative method with binary results ("Positive", "Negative") are summarized in a 2 × 2 contingency table. In this case a = True Positives (TP), b = False Positives (FP), c = False Negatives (FN) and d = True Negatives (TN).

Candidate method	Reference Method/Control Sample		
	Positive	Negative	Total
Positive	a	b	a + b
Negative	c	d	c + d
Total	a + c	b + d	n

where p_1 is the proportion of positive responses for the positive control (and negative responses for the negative control) and $p_2 = 1 - p_1$. Finally, precision is reported as the mean CU% for the two control samples.

Target: Target values for analytical performance should always be based on the Milan hierarchy [27]. Hence, the effect of analytical performance on clinical outcomes takes priority over performance specifications based on biological variation, which in turn takes priority over specifications based on state-of-the art. As the first two criteria cannot be readily applied to qualitative assays, targets are defined in terms of the highest level of analytical performance technically achievable by field methods. These can be derived e.g. from external quality assessment programs [28].

When following a protocol, such as the one described earlier, a total of 20 measurements are performed (10 negative, 5 positive and 5 weak-positive). If discordant results should occur, then those are most likely to be caused by one or more of the 5 weak-positive samples. In this scenario a single incorrect measurement would result in a *CU*% of 9.0%. If the total number of measurements is increased to 40 then two incorrect results would lead to a *CU*% of 9.5%. It therefore seems reasonable, to set the target for precision to ≤10%.

Accuracy: For binary testing systems, accuracy is defined in terms of how well the system is able to correctly identify a condition. Thus, accuracy is the proportion of true results (true-positive and true-negative) in the population tested. It can be expressed as total accuracy, true-positive rate (TPR), false-positive rate (FPR), true-negative rate (TNR) or false-negative rate (FNR):

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (4)$$

$$TPR = \frac{TP}{(TP + FN)} = 1 - FNR \quad (5)$$

$$FPR = \frac{FP}{(FP + TN)} = 1 - TNR \quad (6)$$

$$TNR = \frac{TN}{(TN + FP)} = 1 - FPR \quad (7)$$

$$FNR = \frac{FN}{(FN + TP)} = 1 - TPR \quad (8)$$

with TP = True-Positives, TN = True-Negatives, FP = False-Positives and FN = False-Negatives.

When constructing confidence intervals, these should be calculated as *score confidence intervals*, as described by Agresti and Coull [1,29].

Target: For reasons similar to those discussed earlier, performance specifications for accuracy are usually defined in terms of state-of-the-art. Once again, it appears reasonable to accept one incorrect result in a small set of 20 measurements, which yields a total agreement of 95%. The target for accuracy should therefore be set to ≥95%.

It is important to note that TPR, FPR, TNR and FNR do not represent all aspects of a method's performance, as they use only part of the information of the contingency table. For this reason, a number of metrics have been described, to better incorporate all measures pertaining to diagnostic performance. These include, among others, the F1-score [30] and the Matthews Correlation Coefficient (MCC) [31]. The latter takes benefit of TP, TN, FP and FN results and is therefore a robust indicator of agreement, which can also be applied to unbalanced data. It is calculated as follows [32,33]:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN)(TN + FP)(TP + FP)(TN + FN)}} \quad (9)$$

Target: The values can range from −1 to 1, with −1 indicating total disagreement between prediction and true values, 0 representing random distribution and 1 indicating perfect agreement between prediction and true values. The *MCC* reaches 0.60 when 80% of cases are correctly predicted and 0.80 when 90% of cases are correctly predicted [32]. However, when evaluating accuracy with control samples, a total accuracy of 95% should be achieved, which corresponds to an *MCC* of 0.90 (Table 2).

Table 2
Summary statistics for precision & accuracy testing of a qualitative drug-of-abuse screening assay. Results are summarized in a 2 × 2 table, and symmetry testing is performed with the help of McNemar's test. Accuracy is expressed as Total Accuracy, TPR, FPR, TNR, FNR and with the help of the Matthews Correlation Coefficient.Test precision is smaller than target precision and thus acceptable. The total accuracy is larger than the target and therefore also acceptable. A value for the Mathews correlation coefficient ≥0.90 indicates that > 95% of cases were predicted correctly.

Statistics				
Test (Result)		Reference (Control Sample / Diagnosis)		
		Positive	Negative	Total
	Positive	22	1	23
	Negative	0	17	17
	Total	22	18	40
McNemar statistic		0		
p-Value (Exact binomial)		1.0000 (α = 0.05)		
Symmetry Test Pass/Fail?		Pass		
Precision Estimates				
Precision (CU%)*		5.2%		
Target Precision		10.0%		
Pass/Fail?		Pass		
Bias Estimates				
Accuracy		97.5%		
True-positive rate (TPR)		100.0%		
False-positive rate (FPR)		5.6%		
True-negative rate (TNR)		94.4%		
False-negative rate (FNR)		0.0%		
Matthews Correlation Coefficient		0.95		
Target Accuracy		95.0%		
Pass/Fail?		Pass		

* CU% = Coefficient of Unalikeability.

Table 3
Comparison data between methods with ordinal results are summarized in a $k \times k$ contingency table, which is an expansion of the 2 × 2 table to more than two categories.

Candidate Method	Reference Method / Control Sample				
	1	2	…	k	Total
1	p_{11}	p_{12}	…	p_{1k}	$p_{1\cdot}$
2	p_{21}	p_{22}	…	p_{2k}	$p_{2\cdot}$
⋮	⋮	⋮	⋱	⋮	⋮
k	p_{k1}	p_{k2}	…	p_{kk}	$p_{k\cdot}$
Total	$p_{\cdot 1}$	$p_{\cdot 2}$	…	$p_{\cdot k}$	1

2.1.2. Semi-quantitative assays

Semi-quantitative assays represent an extension of qualitative assays to more than two categories. Thus, the experimental design discussed for qualitative assays applies here as well. At least 10 to 20 control samples should be measured per category, with at least 30 measurements in total [4]. Results are summarized in a $k \times k$ contingency table (Table 3) or presented in the form of a bubble chart (Fig. 1). This is similar to a scatter plot, in which the size of each bubble is directly related to the number of observations in the group represented by the bubble. Cases with perfect agreement result in bubbles lying on the diagonal, while cases with less than perfect agreement result in bubbles on either side of the diagonal (See Table 4).

Symmetry testing: Once again, the data must be examined for a systematic difference between the results obtained and the target values for the samples. As McNemar's test can be applied to only two categories, it cannot be used for semi-quantitative assays. Rather, Bowker's

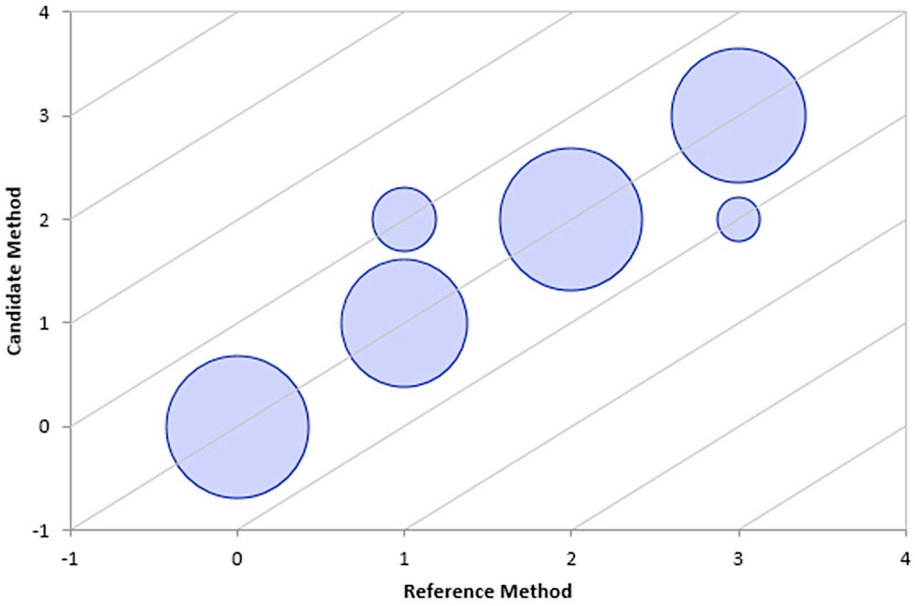

Fig. 1. Bubble chart, depicting the results for a method comparison study between two urine Glucose test-strip assays. Bubble charts are similar to scatter plots, in which the size of each bubble is directly related to the number of observations in the category denoted by the bubble. The above example depicts measurement results of the candidate method, plotted on the y-axis against results of the reference method plotted on the x-axis. For the precision and accuracy experiment, measurement results would be plotted on the y-axis and target values of the control material or the certified reference material would be plotted on the x-axis.

Table 4
Summary statistics for precision and accuracy testing for a semi-quantitative urine Glucose test-strip assay. Precision is expressed as coefficient of unalikeability (*CU*%). The *CU*% is smaller than the target precision; observed accuracy exceeds the target value. Precision and accuracy are therefore acceptable.

Statistics	
n:	40
Categories:	4
Bowker's Test for internal Symmetry	1.00
p-Value (two-tailed)	0.9856 (α = 0.05)
Symmetry Test Pass/Fail?	Pass
Precision Estimates	
Precision (CU%)*	4.5%
Target Precision	6.0%
Pass / Fail?	Pass
Bias Estimates	
Accuracy	97.5%
Target Accuracy	95.0%
Pass/Fail?	Pass

* CU% = Coefficient of Unalikeability.

test for internal symmetry is used. This is identical to McNemar's test, but is used with square tables having more than two categories [34,35]. Similar to McNemar's test, it also investigates the null hypothesis that the cell proportions (p_{ij} and p_{ji}) are symmetric for all pairs of table cells. A statistically significant value (e.g. $p < 0.05$) indicates a substantial difference in the number of observations between the reference method and the candidate method in one or more categories and will need to be investigated further. Alternatives to Bowker's test are tests for marginal homogeneity, such as the Stuart Maxwell test [36] and the more powerful Bhapkar test [37]. The disadvantage of these tests is that they only test for marginal homogeneity, and while symmetry in a contingency table implies marginal homogeneity, the reverse is not true [38]. Bowker's test, on the other hand, performs well with symmetry testing and is simple to compute, which is why it is recommended here.

Precision: Precision is calculated in a similar fashion to qualitative assays, using the coefficient of unalikeability. Calculations are performed separately for each control sample and the total *CU*% is reported as the mean *CU*% of all *k* levels:

$$CU\%_k = 1 - \sum_{i=1}^{n} p_i^2 \tag{10}$$

$$CU\%_{total} = \frac{\sum_{i=1}^{k} CU\%_i}{k} \tag{11}$$

where

$$p_i = \frac{k_i}{n} \tag{12}$$

Target: Target values for performance specifications are again based on state-of-the-art. With a minimal protocol, such as described above, 10 repeat measurements are carried out for each of three categories. In this scenario, one discrepant result leads to a *CU*% of 6.0%. With four categories, a *CU*% of 4.5% is achieved with one discordant result and 9.0% with two discordant results. It is therefore feasible to set the desirable target for precision to ≤6.0% and the acceptable target to ≤9.0%.

Accuracy: For semi-quantitative testing systems, accuracy is defined in terms of how well the system is able to correctly identify the measurement category in question. It is calculated as the proportion of correct results for each category (p_i), divided by the total number of categories (k), multiplied by 100:

$$Accuracy = \frac{\sum_{i=1}^{k} p_i}{k} \times 100 \tag{13}$$

Target: With 10 repeat measurements per category, one incorrect result leads to a total accuracy of 96.7% with three categories (e.g. negative, weak-positive, positive) and 97.5% with four categories (e.g. trace, small, moderate, large). Two incorrect results with 10 repeat measurements and four categories result in a total precision of 95%. Following the "state-of-the-art" principle again, a target accuracy of ≥95% seems reasonable.

2.2. Method comparison

2.2.1. Qualitative assays

The method comparison experiment is carried out with patient samples, reference materials or proficiency testing materials. CLSI recommends to continue testing, until at least 50 positive specimens are obtained with both the test and comparative method. Furthermore, samples with concentrations around the cutoff (e.g. ± 20%) should also be included, to evaluate the performance of the candidate method at this critical concentration [39]. Other authors suggest, that analyzing 40 samples (10 positive, 10 weak positive and 20 negative samples) is sufficient for routine clinical testing [40]. If samples for testing are not

readily obtainable, the method comparison experiment may be carried out with as little as 10 positive and 10 negative specimens [4]. Results are summarized in a 2 × 2 contingency table, which serves as the basis for the calculations. The calculated estimates are termed “Percent Positive Agreement (PPA)” and “Percent Negative Agreement (PNC)”, rather than “sensitivity” and “specificity”, which emphasizes the fact, that these estimates reflect agreement between methods and not diagnostic accuracy.

Formal testing of the degree of agreement between the two methods is carried out with Cohen's kappa (κ) [41]. For binary ratings, i.e. assays with only “positive” or “negative” results, simple *unweighted* kappa is appropriate. The test focuses on the major diagonal of the contingency table, which contains the proportions of agreement between the two tests, and assesses if diagonal counts are significantly larger than those expected by chance alone. It therefore represents the agreement adjusted for that expected by chance and is calculated as the difference between observed agreement (p_o) and agreement expected by chance alone (p_e), divided by the maximum possible difference:

$$\kappa = \frac{(p_o - p_e)}{(1 - p_e)} \tag{14}$$

$$p_o = \sum_{i=1}^{k} p_{ii} \tag{15}$$

$$p_e = \sum_{i=1}^{k} p_{i.} p_{i} \tag{16}$$

with p_{ii}= proportions lying on the main diagonal of the $k \times k$ contingency table (i.e. the proportion of results with perfect agreement), $pi.$ = the sum of proportions in row_i and $p.i$= the sum of proportions in $column_i$.

As the maximum value for p_o is 1, which happens when there is perfect agreement between the two assays, the maximum possible difference for $p_o - p_e$ is $1 - p_e$. The limits of κ are +1.00 (perfect agreement) and −1.00, with values between 0 and −1.00 indicating a level of agreement that would be expected purely by chance. Various interpretations for κ are found in the literature. In the clinical laboratory, the levels of agreement described by Landis and Koch are used most commonly [42]:

Table 5
Summary statistics for a method comparison study between two qualitative drug-of-abuse screening tests. Results are summarized in a 2 × 2 contingency table. Symmetry testing is performed with McNemar's test. A p-value > α indicates acceptable symmetry. The sum of the discordant cells is 3 + 1 = 4 and therefore < 10, which is why hypothesis testing was carried out with an exact binomial test. Agreement between the two methods is statistically evaluated with Cohen's kappa (*unweighted*). The value for Cohen's kappa exceeds the target value and therefore indicates acceptable agreement between the two methods.

Statistics				
Test (Candidate Method)		Reference (Reference Method)		
		Positive	Negative	Total
	Positive	20	3	23
	Negative	1	16	17
	Total	21	19	40
McNemar statistic:		0.25		
p-Value (Exact, binomial):		0.6250 (α = 0.05)		
Symmetry Test Pass / Fail?		Pass		
Agreement		90.0% (76.9% to 96.0%)		
Pos. Agreement		95.2% (77.3% to 99.2%)		
Neg. Agreement		84.2% (62.4% to 94.5%)		
Cohen's kappa		0.80 (0.61 to 0.99)		
p-Value (one-tailed)		< 0.0001		
Target kappa		0.70		
Pass/Fail?		Pass		

Table 6
Linear weights for a 4-category scale.

Candidate Method	Reference Method / Control Sample			
	0	1	2	3
0	1,00	0,67	0,33	0,00
1	0,67	1,00	0,67	0,33
2	0,33	0,67	1,00	0,67
3	0,00	0,33	0,67	1,00

$\kappa \leq 0.20$	Poor agreement
$0.21 \leq \kappa \leq 0.40$	Fair agreement
$0.41 \leq \kappa \leq 0.60$	Moderate agreement
$0.61 \leq \kappa \leq 0.80$	Good agreement
$0.81 \leq \kappa \leq 1.00$	Almost perfect agreement

A high value for kappa indicates strong agreement between measurement results and target values and thus good accuracy. The kappa statistic is dependent on the prevalence of the disease, or in the case of method comparisons, on the proportion of positive samples. Thus, for two method comparison studies with the same proportion of observed agreement, kappa would be highest in the experiment where the proportion of positive samples is closer to 50%. If the proportion of positive samples is very low or very high, the kappa statistic becomes smaller. Caution must therefore be exercised, when selecting samples for qualitative method comparisons, paying close attention to a balanced design (comparable number of positive and negative samples). A sensitivity and specificity of 90% result in a κ = 0.80, which means that out of all expected disagreement by chance, non-random agreement was obtained with the tested method in 80% of cases. A sensitivity and specificity of 80% result in κ = 0.60. Therefore, most sources agree that κ should be > 0.60 and optimally > 0.80, which means that more than half of the non-random cases are classified correctly [3,43]. At least 20, better 30 samples (ideally 10 to 15 positive and 10 to 15 negative samples) should be analyzed, to obtain a reliable value for kappa.

Although Cohen's kappa is the preferred measure of agreement for qualitative and semi-quantitative assays, there are many others. A simple approach is to calculate the overall agreement (“Percent Total Agreement – PTA”), (Table 1):

$$PTA = \frac{(a + d)}{n} \times 100 \tag{17}$$

This does not, however, adequately characterize the agreement between two methods, as ($b + c$) can be identical for two different 2 × 2 tables with very different individual values for b and c. It is

Table 7
Summary statistics of a method comparison study between two semi-quantitative urine Glucose test-strip assays. Symmetry testing is performed with Bowker's test for internal symmetry. A p-value > α indicates acceptable symmetry. Method agreement is evaluated with *weighted* kappa. A one-tailed p-value for kappa, which is < α indicates that kappa differs from 0 and differences detected by kappa are not random. The value for weighted kappa is larger than the target, indicating acceptable agreement between the two methods.

Statistics	
n	40
Levels	4
Bowker's Test for Internal Symmetry:	2.00
p-Value (2-Sided)	0.9197 (α = 0.05)
Symmetry Test Pass/Fail?	Pass
Cohen's Kappa (weighted)	0.92 (0.84 to 1.00)
p-Value (One-sided)	< 0.0001
Target kappa	0.80
Pass/Fail?	Pass

therefore more useful to calculate a pair of agreement measures, Percent Positive Agreement ("PPA") and Percent Negative Agreement ("PNA") [1]:

$$PPA = \frac{a}{(a + c)} \times 100 \quad (18)$$

$$PNA = \frac{d}{(b + d)} \times 100 \quad (19)$$

Again, samples should be carefully selected, ensuring a balanced design, with a comparable proportion of positive and negative samples. The acceptance criteria for PC and NC are 90%, respectively, and should not be lower than 80% (Table 5) [3].

2.2.2. Semi-quantitative assays

The method comparison experiment is carried out with patient samples, reference materials or proficiency testing materials, which should span the entire measurement range (i.e. all categories detected by the assay). At least 10 samples should be included per category, with no less than 30 samples in total [4]. Results are summarized in a $k \times k$ table (Table 3) or depicted graphically in the form of a bubble chart (Fig. 1). With semi-quantitative assays three or more ordered categories are present. Therefore the extent of disagreement can vary (e.g. one, two or more categories), and this must also be taken into account, when calculating kappa. Ordinary unweighted kappa, which is used for binary systems, is therefore not appropriate in this case and weighted kappa must be used instead. Weights are assigned to non-diagonal cells of the contingency table, depending on their distance from the diagonal. The weight is directly related to the distance of the cell from the diagonal and becomes smaller, as the distance increases. Two weighting systems are described: linear and quadratic weighting. While with linear weighting, weights remain constant, irrespective of the magnitude of the difference between the two results, with quadratic weighting it increases with increasing difference. For routine laboratory testing, linear weighting is employed:

$$w_{ij} = 1 - \frac{|i - j|}{(k - 1)} \quad (20)$$

with $i =$ row, $j =$ column, $|i - j| =$ the number of categories of disagreement and $k =$ total number of categories [44]. On a 4-category scale, a disagreement by one category would therefore be weighted with 0.67, a difference by two categories with 0 and so on (Table 6).

Target: Specifications for κ_w are somewhat tighter than for simple unweighted kappa. Ideally κ_w should be ≥ 0.90 and minimally ≥ 0.70 (Table 7) [3].

Although weighted kappa is the preferred measure of agreement for semi-quantitative assays, an alternative approach is suggested for immunological assays in virus diagnostics by Rabenau et al. The authors propose, that the final evaluation of interchangeability of the candidate and reference methods is judged according to the overall extent of disagreement, with discrepancies of up to ± 1 titer levels being acceptable [39].

3. Conclusion

Qualitative and semi-quantitative methods are common in the clinical laboratory. DIN ISO 15189, DIN EN ISO 17025 and CAP require that the performance of these methods is verified prior to their implementation for reporting patient results. If numeric values are generated in the form of measurement signals during the analytical process, than these can be used to assess precision and accuracy in an analogous fashion to quantitative assays. For qualitative assays with strict binary results, appropriate calculations for precision and accuracy must be used. These can include symmetry testing with McNemar's test, precision testing with the coefficient of unalikeability and expressing accuracy in terms of total accuracy, true- and false positive, as well as true- and false-negative results. Finally, the Matthews Correlation Coefficient provides a reliable indication of agreement between predicted and true values. Method comparison experiments can be carried out with as few as 20 patient samples and agreement between methods is expressed with the help of Cohen's kappa.

For semi-quantitative assays with ordinal results precision and accuracy can be estimated analogously to qualitative assays by extending the appropriate calculations to more than two categories. Symmetry testing is performed with Bowker's test for internal symmetry and weighted kappa is used to quantify agreement between methods. For immunological assays in virus diagnostics titer levels may be used.

All statistical graphs and tables were created with Abacus 2.0 (LABanalytics GmbH) [45].

References

[1] CLSI, User Protocol for Evaluation of Qualitative Test Performance; Approved Guideline - Second Edition. CLSI Document EP12-A2, Clinical and Laboratory Standards Institute, Wayne, PA, 2008.
[2] N. Gunnar, Before defining performance criteria we must agree on what a qualitative test procedure is, Clin. Chem. Lab. Med. 53 (2015) 939–941.
[3] T. Kouri, P. Laippala, D. Kutter, et al., Quality specifications for ordinal scale measurements with multiproperty (multiple) urine test strips, Scand. J. Clin. Lab. Invest. 59 (1999) 523–526.
[4] E. Topic, N. Nikolac, M. Panteghini, et al., How to assess the quality of your analytical method? Clin. Chem. Lab. Med. 53 (2015) 1707–1718.
[5] P.H. Petersen, N.G. Christensen, S. Sandberg, et al., How to deal with semi-quantitative tests? Application of an ordinal scale model to measurements of urine glucose, Scand. J. Clin. Lab. Invest. 69 (2009) 662–672.
[6] DIN EN ISO 15189, Medical Laboratories - Particular Requirements for Quality and Competence, International Organization for Standardization, Geneva, Switzerland, 2014.
[7] DIN EN ISO/IEC 17025, General Requirements for the Competence of Testing and Calibration Laboratories, International Organization for Standardization, Geneva, Switzerland, 2017.
[8] CAP - All Common Checklist, College of American Pathologists, 2015.
[9] S.J. Sarewitz, CAP Accreditation Programs Audioconferences and Webinars CAP Accreditation Requirements for Validation of Laboratory Tests, http://webapps.cap.org/apps/docs/education/lapaudio/pdf/011613.presentation.pdf, (2013) (Accessed: 15 May, 2019).
[10] P. Pereira, Basic Concepts on Validation of Qualitative Tests, https://www.westgard.com/validating-qualitative-tests.htm (2018) (Accessed: December 29, 2018).
[11] ISO 3534-1:2006, Statistics — Vocabulary and symbols — Part 1: General Statistical Terms and Terms Used in Probability, International Organization for Standardization, Geneva, Switzerland, 2006.
[12] National Association of Testing Authorities, Guidelines for the Validation and Verification of Quantitative and Qualitative Test Methods - Technical Note 17 (Australia), (2012).
[13] ISO 16140, Microbiology of Food and Animal Feeding Stuffs—Protocol for the Validation of Alternative Methods. Annex M. Interlaboratory Study of Qualitative Methods: Criteria of Accordance, Concordance and Concordance Odds Ratio, International Organization for Standardization, Geneva, Switzerland, 2003.
[14] S.D. Langton, R. Chevennement, N. Nagelkeke, et al., Analysing collaborative trials for qualitative microbiological methods: accordance and concordance, Int. J. Food Microbiol. 79 (2002) 171–181.
[15] H. van der Voet, W.D. van Raamsdonk, Estimation of accordance and concordance in inter-laboratory trials of analytical methods with qualitative results, Int. J. Food Microbiol. 95 (2004) 231–234.
[16] J. Mandel, Repeatability and reproducibility for pass/fail data, J. Test. Eval. 25 (1997) 151–153.
[17] P. Wilrich, The determination of precision of qualitative measurement methods by interlaboratory experiments, Accred. Qual. Assur. 15 (2010) 439–444.
[18] G. Kader, M. Perry, Variability for categorical variables, J. Stat. Ed. 15 (2007) 1–16.
[19] M. Perry, G. Kader, Variation as unalikeability, Teach. Stat. 27 (2005) 58–60.
[20] T. Gordon, Is the standard deviation tied to the mean? Teach. Stat. 8 (1986) 40–42.
[21] E. Theodorsson, B. Magnusson, I. Leito, Bias in clinical chemistry, Bioanalysis 6 (2014) 2855–2875.
[22] K. Linnet, K.G.M. Moons, J.C. Boyd, Chapter 2, analytical and clinical evaluation of methods, in: N. Rifai, A.R. Horvath, C.T.T. Wittwer (Eds.), Tietz Fundamentals of Clinical Chemistry and Molecular Diagnostics, 8th ed., Elsevier/Saunders, St. Louis, 2019.
[23] CLIA, 42 CFR 493.1253 - Standard: Establishment and Verification of Performance Specifications, (2003).
[24] D.M. Wolk, Statistics for method verification of qualitative assays in clinical microbiology, Clin. Microbiol. Newsl. 39 (2017) 59–68.
[25] D.J. Sheskin, Test 16, the Chi-Square test for $r \times c$ tables (nonparametric test emplyed with categorical/nominal data), in: D.J. Sheskin (Ed.), Handbook of Parametric and Nonparametrich Statistical Procedures, 5th ed., Taylor & Francis Ltd, Boca Raton, 2011.
[26] O.A. Adedokun, W.D. Burgess, Analysis of paired dichotomous data: a gentle

introduction to the McNemar test in SPSS, J. MultiDisc. Eval. 8 (2012) 125–131.
[27] S. Sandberg, C.G. Fraser, A.R. Horvath, et al., Defining analytical performance specifications: Consensus Statement from the 1st Strategic Conference of the European Federation of Clinical Chemistry and Laboratory Medicine, Clin. Chem. Lab. Med. 53 (2015) 833–835.
[28] F. Ceriotti, P. Fernandez-Calle, G.G. Klee, et al., On behalf of the EFLM task and finish group on allocation of laboratory tests to different models for performance specifications (TFG-DM), criteria for assigning laboratory measurands to models for analytical performance specifications defined in the 1st EFLM Strategic Conference, Clin. Chem. Lab. Med. 55 (2017) 189–194.
[29] A. Agresti, B. Coull, Approximate is better than exact for interval estimation of binomial proportions, Am. Stat. 52 (1998) 119–126.
[30] Q. Zou, S. Xie, Z. Lin, et al., Finding the best classification threshold in imbalanced classification, Big. Data Res. 5 (2016) 2–8.
[31] B.W. Matthews, Comparison of the predicted and observed secondary structure of T4 phage lysozyme, Biochim. Biophys. Acta - Protein Struct. 405 (1975) 442–451.
[32] M. Vihinen, et al., BMC Genomics 13 (Suppl. 4) (2012) 1–10 S2.
[33] S. Boughorbel, F. Jarray, M. El-Anbari, Optimal classifier for imbalanced data using Matthews correlation coefficient metric, PLoS One 12 (2017) e0177678.
[34] A.H. Bowker, A test for symmetry in contingency tables, J. Am. Stat. Assoc. 43 (1948) 572–574.
[35] P.R. Yarnold, UniODA vs. Bowker's test for symmetry: diagnosis before vs. after treatment, Optim. Data Anal. 4 (2015) 29–31.
[36] A. Stuart, A test for homogeneity of the marginal distributions in a two-way classification, Biometrika 42 (1955) 412–416.
[37] V. Bhapkar, A note on the equivalence of two test criteria for hypotheses in categorical data, Am. Stat. Assoc. 61 (1966) 228–235.
[38] D. Rahardja, Y. Yang, Z. Zhang, A comprehensive review of the two-sample independent or paired binary data, with or without stratum effects, J. Mod. Appl. Stat. Methods 15 (2016) 215–223.
[39] M. Kortenbusch, A. Berger, H.F. Rabenau, Muster VA aus dem Institut für Medizinische Virologie Universitätsklinikum Frankfurt, https://www.g-f-v.org, (2019) Muster Verfahrensanweisung Validierung – aktualisiert mit Änderungen - final 05.09.2017.pdf (Accessed: 15 June, 2019).
[40] H.F. Rabenau, H.H. Kessler, M. Kortenbusch, et al., Verification and validation of diagnostic laboratory tests in clinical virology, J. Clin. Virol. 40 (2007) 93–98.
[41] J. Fleiss, Statistical Methods for Rates and Proportions, 3rd edition ed., Wiley & Sons, New York, 1981.
[42] J. Landis, G. Koch, The measurement of observer agreement for categorical data, Biometrics 33 (1977) 159–174.
[43] M. Miler, Verification of Qualitative Methods, in: 15th EFLM Continuing Postgraduate Course in Clinical Chemistry and Laboratory Medicine: How to Assess the Quality of Your Method? https://www.em.eu/files/efcc/Zagreb-Miler.pdf, (2015) (Accessed: 15 July, 2019).
[44] J. Sim, C.C. Wright, The kappa statistic in reliability studies: use, interpretation, and sample size requirements, Phys. Ther. 85 (2005) 257–268.
[45] LABanalytics GmbH, Abacus 2.0 - Method-Validation, Quality-Control and Statistics Add-in for Microsoft Excel, https://www.lab-analytics.com, (2019).