PUBLISHED BY

INTECH
open science | open minds

World's largest Science,
Technology & Medicine
Open Access book publisher

3,350+
OPEN ACCESS BOOKS

108,000+
INTERNATIONAL
AUTHORS AND EDITORS

115+ MILLION
DOWNLOADS

BOOKS
DELIVERED TO
151 COUNTRIES

AUTHORS AMONG
TOP 1%
MOST CITED SCIENTIST

12.2%
AUTHORS AND EDITORS
FROM TOP 500 UNIVERSITIES

CLARIVATE ANALYTICS
BOOK CITATION INDEX
INDEXED

WEB OF SCIENCE™

Selection of our books indexed in the
Book Citation Index in Web of Science™
Core Collection (BKCI)

Chapter from the book *Proof and Concepts in Rapid Diagnostic Tests and Technologies*
Downloaded from: http://www.intechopen.com/books/proof-and-concepts-in-rapid-diagnostic-tests-and-technologies

Interested in publishing with IntechOpen?
Contact us at book.department@intechopen.com

Chapter 8

Evaluation of Rapid Diagnostic Test Performance

Paulo Pereira

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/64179

Abstract

Rapid diagnostic tests are used for the determination of binary qualitative results not only uniquely in nonhospital-based but also in hospital-based tests. Principally, in developing countries, rapid diagnostic tests are the primary option since tests to be used in medical laboratories are discarded due to the higher cost. The test's performance is evaluated to assure that the chance of results to be false is clinically acceptable. Therefore, the diagnostic accuracy of results (diagnostic sensitivity and specificity) is assessed to guarantee the safety of postclinical decision. The statistical approach requires that representative samplings of the populations of infected and healthy individuals are tested. The area under the receiver-operating characteristic (ROC) curve is a complementary measurement using the same samplings. It represents the diagnostic accuracy in a single outcome. When samplings with known diagnostics are unavailable, samplings with known outcomes from a comparative test are used to determine the agreement of results. However, this approach is secondary, due to diagnostic accuracy to be unmeasurable. The seronegative period is another critical measurement that allows determining an individual biological bias during a period where results of an infected individual are false-negatives due to seroconversion. The claimed requirements should be defined for diagnostic accuracy and agreement outcomes. A spreadsheet is used to estimate the results considering the absolute value and the 95% confidence interval.

Keywords: clinical decision, point-of-care testing, quality assurance, quality control, rapid diagnostic test

1. Introduction

Point-of-care testing (POCT) is defined by the International Organization for Standardization (ISO) as the "testing that is performed near or at the site of the patient with the result leading to

INTECH
open science | open minds

© 2016 The Author(s). Licensee InTech. This chapter is distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

a possible change in the care of the patient" [1]. Clinical and Laboratory Standards Institute (CLSI) recognizes this definition and proposes another: "testing performed outside a central laboratory environment, generally nearer to, or at the site of the patient/client" [2]. Currently, 31% of the global *in vitro* diagnostics' (IVD) commercial tests are POCT [3]. The US POCTs represent close to 45% of the IVD market, followed by Europe and Japan. The global POCT market is estimated to increase to reach US$ 30 billion by 2020, with a compound annual growth rate close to 10% [4]. The POCT for the detection of infectious disease agents is 21.9% of the forecast POCT global market segments in 2016 [5].

Rapid diagnostic test (RDT) is one type of POCT intended to a fast detection of the antigens or antibodies present in human samples. RDT is a qualitative (or semiquantitative) test, and it is defined as the test "that detect and/or identify a particular analyte, constituent, or condition (note: this term applies to tests that detect whether a particular analyte, constituent, or condition is present or absent and sometimes assigned a positive degree (i.e., 1+, 2+)" (entry 4 in Ref. [2]). According to the International Vocabulary of Metrology (VIM), generally RDT result is analogous to nominal quantity (entry 4.26 of Ref. [6]) tests. This quantity is defined as the "rounded or approximate value of a characterizing amount of a measuring instrument or measuring system that guides its appropriate use," for example, positive/negative. Nominal quantities are sometimes confused with ordinal quantities (entry 1.26 of Ref. [6]), where a value's classification is according to a decision value on an ordinal scale, for example, a result equal or higher than the clinical decision point ("cutoff") is classified as positive, and if below —negative. RDTs are usually used as screening tests, namely, in cases where the cost per result must be lower than that if it is used in the medical laboratory test such as that happens predominantly in developing countries. They are also used when the need for an outcome is an emergent such as in an emergency room. RDT should provide accurate, precise, and reliable test results. Unreliable results have a significant probability to be untrue, affecting the clinical decision directly. According to the current good laboratory practice, RDT results should not be final results. However, in developing countries, it is frequently considered an end outcome, for what the physician should understand the risk of clinical decisions' failure due to false results. Accordingly, the laboratorian should perform a set of quality control tests to evaluate the accuracy of the RDT results. This is intended to determine and verify that the error associated to the *in vivo* results is not clinically significant (allowable error), that is, it does not affect the clinical decision significantly. When quality control tests are not performed, the residual risk (entry 2.29 of Ref. [7]) linked to the chance of incorrect clinical decisions due to abnormal results is unknown. Evidenced-based reported cases [8–11] demonstrate the impact of bad quality control practices of POCT in the clinical decisions, contributing to a discussion on the accuracy of POCT results and its association with unsafe clinical decisions [12].

Quality control (entry 3.3.7 of Ref. [13]) is a part of the quality assurance (entry 3.3.6 of Ref. [13]). It is a set of models designed to monitor the test's results. Its goal is to assure that the outcomes (true result + error) are not significantly distant from the *in vivo* results. Test evaluation should be performed determining the diagnostic accuracy of the assay focused on diagnostic sensitivity and specificity. The area under the receiver-operating characteristic curve (AUC) is a complementary measurement that permits the calculus of diagnostic

uncertainty in a single outcome. Both determinations are statistical measures of the performance of a binary result (positive or negative). However, to use these applications, there must be available samplings from infected individuals D_1 and healthy individuals D_0. When the comparator is other than diagnostic, it should be determined as the agreement of results. The seronegative window period is another required measurement in an RDT evaluation. This model evaluates the seronegative period in an infected individual, during which the positive results are biologically biased, that is, they are systematically false negative.

Associated with all the presented approaches is the absence of metrological traceability of the results due to the unavailability of certified reference materials or reference methods (entry 2.41 of Ref. [6]). Therefore, RDTs are classified as untraceable tests. For a thorough discussion of metrological traceability in medical laboratories' tests, further information can be found elsewhere [14].

This chapter reviews and discusses models of diagnostic accuracy, agreement of results, and seronegative window period applied to the RDTs' evaluation. It is presented as an example a case study considering the worst scenario in a blood establishment: the use of a small number of samples where the evaluation cost must be minimum. Despite this example contributes to a less reliable estimation, its role in the consistency of the assessment remains significant. Differently from the tests' evaluation, the determination of measurement uncertainty (entry 2.26 of Ref. [3]) is not systematically applied to RDTs. However, it is considered in the estimates, considering risk-based thinking principles [15–17]. A flowchart summarizes and hierarchizes the alternatives in the selection of evaluation models. Internal quality control (entry 3.3 of Ref. [18]) and external quality assessment/proficiency testing (entry 5.1 of Ref. [19]) models are out of the chapter's scope. Ethical issues are also out of the scope. A report on legal and ethical questions of RDT is found elsewhere [20].

The theoretical principles were applied using standard spreadsheet software (Microsoft® Excel®), freely available at DOI: 10.13140/RG.2.1.1123.1766.

2. Comparison of results

2.1. Diagnostic accuracy I: 2 × 2 contingency tables

2.1.1. General concept

Diagnostic accuracy models are used to determine the RDT performance using diagnostic accuracy criteria (i.e., disease and nondisease), adopting a Bayesian probability framework. This is a concept of probability where the hypothesis of a certain condition (e.g., infected and healthy) is measured in a sampling. For this measurement, two samplings are required: one featuring infected individuals and another featuring healthy individuals. Both samples should be carefully selected to minimize the probability of sources of biological bias considerably, principally "spectrum bias." Pepe [21] defined this type of bias as the "bias between estimated test performance and true test performance when the sample used for evaluating an assay does not properly represent the entire disease spectrum over the target (intended-use)

population" (entry 4.2 of Ref. [21]). The evaluator of test report recognizes there is always an estimated risk of biased results, mainly from samples of individuals in the seroconversion window period (see Section 2.4.), where the concentration of antibodies or antigens is below the RDT "cutoff." **Figure 1** illustrates the case when results of infected and healthy individuals' samples are misclassified, expressing false binary results. In this example, the results are incorrectly equal, higher or lower than the "cutoff." RDTs use a visual "cutoff" point, instead of a numerical value on an ordinal scale such as that used in diagnostic immunoassays.

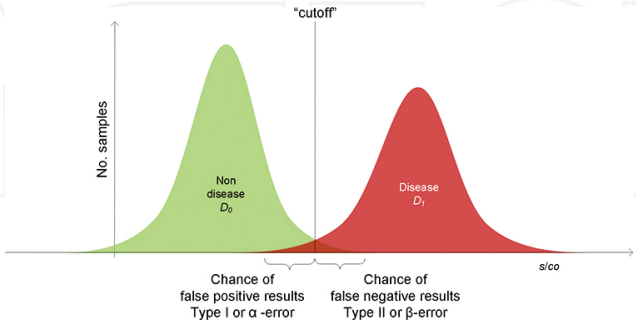

Figure 1. Classification of false results (biased results).

The performance of results is measured using diagnostic sensitivity and diagnostic specificity essentially. These terms are also recognized as clinical sensitivity, sensitivity or true positive rate and clinical specificity, specificity or true negative rate, respectively.

2.1.2. Diagnostic sensitivity

Diagnostic sensitivity *Se* [%] measures the percentage of true positive results in an infected individuals' sampling, that is, the true-positive results rate in percentage (entry 5.3 of Ref. [22]). It is determined by the mathematical model (entry 10.1.1 of Ref. [22]):

$$Se[\%] = \frac{TP}{TP + FN} \cdot 100 \tag{1}$$

where *TP* is the total of true-positive results, and *FN* is the total of false-negative results. The RDT should have a high diagnostic sensitivity to assure a nonsignificant residual risk of the erroneous clinical decision. High sensitivity suggests a high chance of true-positive results. The risk of producing false-positive or false-negative results is readily identified by the statistician to the statistical hypothesis testing as β-error or type II error or α-error or type I error, respectively. Diagnostic sensitivity is equal to $1 - FN = 1 - \beta$-error, for what it is similar to the concept of the statistical power of a binary hypothesis test (entry 2.1 of Ref. [23]), representing the capability of an assay to detect a real effect. Detailed information about statistical hypothesis tests and diagnostic accuracy tests is found elsewhere [21, 23].

The false-negative rate *FN* [%] determines the percentage of false-negative results in the infected individuals' sampling. It is referred to as "diagnostic uncertainty," since it gives a primary sense of the risk of uncertain results [24]. Consequently, "diagnostic uncertainty" of a RDT is defined as "the risk of false results." This concept is analogous to the "measurement uncertainty" intended to be applied to numerical quantity results (entry 1.20 of Ref. [6]). The determination of measurement uncertainty is demonstrated uniquely in ordinal diagnostic tests [25, 26]. In such cases, this terminology could be adapted to "measurement uncertainty of binary results." Measurement uncertainty in an RDTs' results is associated principally with the effect of between-users reading in measures of imprecision.

The false-negative rate is determined by the mathematical model (entry 10.1.4 of Ref. [22]):

$$FN[\%] = \frac{FN}{TP + FN} \cdot 100 \tag{2}$$

2.1.3. Diagnostic specificity

Diagnostic specificity *Sp* [%] measures the percentage of true-negative results in a healthy individuals' sampling, that is, the true-negative results rate in percentage (entry 5.3 of Ref. [22]). It is determined by the mathematical model (entry 10.1.1 of Ref. [22]):

$$Sp[\%] = \frac{TN}{FP + TN} \cdot 100 \tag{3}$$

where *TN* is the total of true-negative results and FP is the total of false-positive results. The false-positive rate FP [%] determines the percentage of false-positive results in the healthy individuals' sampling. In the same manner as the *FN* [%], it could be referred as "diagnostic uncertainty" [24]. It is determined by the mathematical model (entry 10.1.4 of Ref. [22]):

$$FP[\%] = \frac{FP}{FP + FN} \cdot 100 \tag{4}$$

The true and false results can be expressed in a 2 × 2 contingency table as displayed in **Table 1**.

Candidate test results	**Diagnostic accuracy criteria**		**Total**
	Positive (disease, $D = 1$)	Negative (nondisease, $D = 0$)	
Positive ($y = 1$)	True-positive results (*TP*)	False-positive results (*FP*) α-error	$TP + FP$
Negative ($y = 0$)	False-negative results (*FN*) β-error	True-negative results (*TN*)	$FN + TN$
Total	$TP + FN$	$FP + TN$	N

Table 1. 2 × 2 contingency table for diagnostic accuracy.

Predominantly, in rare tests, it could be difficult to have samples with an accurate diagnostic, for what comparative test results should be used when the diagnosis is unknown (see Section 2.3).

2.1.4. Inference of the results for the 95% confidence interval

The rates are applied uniquely to the sampling. They cannot be inferred to the populations. A diagnostic sensitivity equal to 100% is interpreted as a 100% probability of a result of an infected individual from the infected individuals' sampling to be a true positive. It cannot be understood as the likelihood of a person from the population of infected subjects to have a true-positive result.

Nonetheless, an inference could be used to the people of the same characteristics of sampling. Commonly, a confidence interval is used to infer to 95% of the population (95% CI), considering a risk of untrueness throughout α or β-error. To determine the 95% score confidence limits, the low limit and high limit are determined. Considering the level of trust equal to $1 - \alpha$-error, the significance level is equal to 5%. In practice, it can be statistically considered that it is 95% confident about the inclusion of true results in the interval or the error margin of 5% false results attributed to that statement.

For the diagnostic sensitivity, the low-limit interval LL_{se} and high-limit interval LL_{se} are computed from the mathematical models (entry 10.1.3 of Ref. [22]):

$$LL_{se}[\%] = \frac{Q_{1,se} - Q_{2,se}}{Q_{3,se}} \cdot 100 \tag{5}$$

$$HL_{se}[\%] = \frac{Q_{1,se} + Q_{2,se}}{Q_{3,se}} \cdot 100 \tag{6}$$

where $Q_{1,se} = 2 \cdot TP + 1.96^2$, $Q_{2,se} = 1.96 \cdot \sqrt{1.96^2 + 4 \cdot TP \cdot FN / (TP + FN)}$, and $Q_{3,se} = 2 \cdot (TP + FN + 1.96^2)$.

Low and high limits for the test's diagnostic specificity are computed, respectively, by (entry 10.1.3 of Ref. [22])

$$LL_{sp}[\%] = \frac{Q_{1,sp} - Q_{2,sp}}{Q_{3,sp}} \cdot 100 \tag{7}$$

$$HL_{sp}[\%] = \frac{Q_{1,sp} - Q_{2,sp}}{Q_{3,sp}} \cdot 100 \tag{8}$$

where $Q_{1,sp} = 2 \cdot TN + 1.96^2$; $Q_{2,se} = 1.96 \cdot \sqrt{1.96^2 + 4 \cdot FP \cdot TN / (FP + TN)}$, and $Q_{3,sp} = 2 \cdot (FP + TN + 1.96^2)$.

2.1.5. Overall accuracy

Efficiency is an overall estimation of the accuracy in the samplings expressing the percentage of true results in the total of results (entry 4 of Ref. [27]). It is determined by the mathematical model (entry 9.1 of Ref. [27]):

$$Eff[\%] = \frac{TP + TN}{N} \cdot 100 \tag{9}$$

2.1.6. Predictive values

The prediction value of negative results, positive results, or both in a sampling or population has interested the clinical field. Physicians may wish to know the probability of true result to be associated with the presence or absence of a disease, and these can be provided easily.

Predictive value of a positive result or positive predictive value *PPV* [%] measures the percentage of infected individuals in the positive results sampling (entry 5.3 of Ref. [22]). It is determined by the mathematical model (entry 10.3.1 of Ref. [22]):

$$PPV[\%] = \frac{TP}{TP + FP} \cdot 100 \tag{10}$$

Predictive value of a negative result or negative predictive value *NPV* [%] measures the percentage of healthy individuals in the negative results sampling (entry 5.3 of Ref. [22]). It is determined by the mathematical model (entry 10.3.1 of Ref. [22]):

$$NPV[\%] = \frac{TN}{TN + FN} \cdot 100 \tag{11}$$

2.1.7. Prevalence of infected individuals' sampling

Prevalence is not an indicator of performance, but an indicator of the frequency of infected individuals' samples in the total number of subjects. Prevalence of infected individuals *Pr* [%] measures the percentage of infected persons in the total of tested subjects (entry 5.3 of Ref. [22]). It is determined by the mathematical model (entry 10.3.1 of Ref. [22]):

$$Pr[\%] = \frac{TP + FN}{N} \cdot 100 \tag{12}$$

2.1.8. Reproducibility conditions of the study

The sample's measurement should occur in reproducibility conditions, to assure the study evaluates the major error components of the test. For example, if the RDT is to be used by more than one user, it should be tested by all, or if unfeasible, by a representative sampling. Usually, it is recommended the evaluation take place in a period during 10–20 days (entry 9.3 of Ref. [22]). This period is understood as the time when all significant causes of error are expected to happen. However, batch-to-batch variation is commonly omitted, which requires that each batch should be evaluated independently. Sometimes, the evaluations of different batches are done by the inference of one batch's evaluation, which is statistically incorrect since a single batch is not representative of a reagent's manufacturer production.

2.1.9. Interference of results

The samples should be verified taking into account the causes of pre-analytical interference to minimize the risk of false results, biasing diagnostic sensitivity and specificity results. The contribution to false results of anticoagulant, bilirubin, erythrocytes, hemoglobin, dialysis [28], disease effects [29], drugs' effects [30], herbs, and natural products is recognized [31]. The laboratorian could identify the documented interference in published literature or the manufacturer's paper inserted in the reagent kit.

2.1.10. Re-evaluation

The RDT re-evaluation should occur principally when significant epidemiological changes happen; however, other changes require a re-evaluation such as the alteration of the users or even a change in the reagent batch.

2.2. Diagnostic accuracy II: area under the receiver operating characteristic curve

2.2.1. Receiver operating characteristic curve

ROC curve could be defined as "a graphical description of test performance representing the relationship between the diagnostic sensitivity and the false-positive rate" (entry 4.2 of Ref. [32]). **Figure 2** shows three ROC curves. ROC A assures 100% true results using infinite number of determinations from samplings of infected and healthy individuals, which is an unrealistic case; ROC B assumes a realistic situation where a negligible probability of false result happens; and ROC C assures 100% true false results, which is an unrealistic example. A curve following closer to the left-hand border and then the top border represents accurate test's results. ROC is purely a graphical plot exhibiting the performance of an assay as the "cutoff" differs. Accordingly, its application to the RDT user is hard to apply, since there is no numerical "cutoff" but uniquely a visual "cutoff," which cannot be expressed as a numerical quantity by the user. A semiquantitative classification such as a positive degree is not applied usually to these tests.

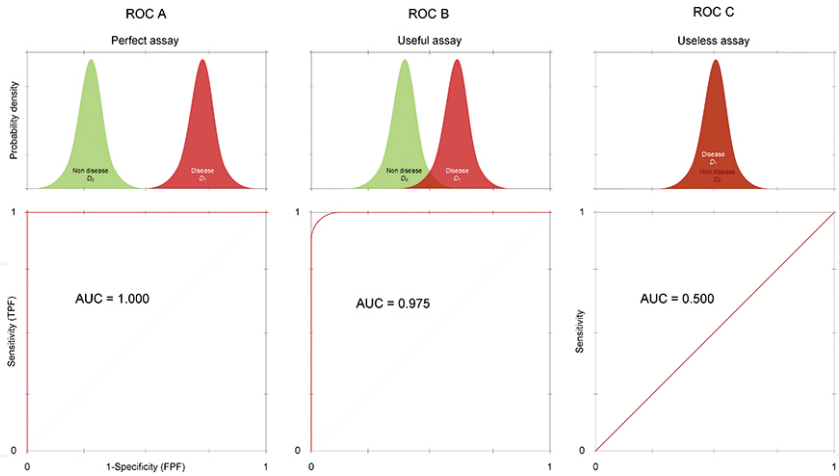

Figure 2. Three hypothetical ROC curves.

2.2.2. Area under the receiver operating characteristic curve

The area under the receiver operating characteristic curve (AUC) determines the ability of a test to classify accurately or discriminate samples. It ranges from 0.5 to 1, signifying respectively, binary results randomly assigned to a sample and a perfect test discrimination. It could be measured using parametric or nonparametric models [24], such as the D'Agostino-Pearson normality test AUC_p [33] and the Mann-Whitney U statistical test AUC_{np} [34], correspondingly. Both models could be computed using the data previously requested in Section 2.1.:

$$AUC_p = FPF(c) \cdot TPF(c) \tag{13}$$

where $FPF(c)$ is the false-positive rate for a defined "cutoff" value c and $TPF(c)$ is the true-positive rate for the same "cutoff" [35].

$$AUC_{np} = U / n_1 \cdot n_0 \tag{14}$$

where $U = (n_1 \cdot n_0 + n_0 \cdot (n_0 + 1))/(2 - R)$, where R is the rank sum of squares; n_1 is the number of negative specimens; and n_0 is the number of positive samples [34].

The difference between one and the AUC could also be referred as "diagnostic uncertainty" of estimation.

Test discrimination could be classified according to the AUC, using the following ratings [36]:

– if $AUC \in [0.50, 0.70]$ the distinction is poor;

– if $AUC \in [0.70, 0.80]$ the distinction is acceptable;

– if $AUC \in [0.80, 0.90]$ the distinction is excellent;

– and if $AUC \in [0.90, 1.00]$ the distinction is outstanding.

The accuracy of AUC is restricted by the number of true and false results. It gives an overall evaluation of RDT performance, complementing diagnostic sensitivity and specificity estimates [24]. As in Section 2.1.4, a confidence interval should be related to the AUC to evaluate if changes in AUC between classifiers are statistically significant in the inferred range. The report evaluator could erroneously infer about the capacity of an RDT to discriminate the results if the absolute values are considered uniquely. The 95% CI could be computed according to the model

$$AUC \pm 1.96 \cdot \mathrm{SE[AUC]} \tag{15}$$

Where SE[AUC] is the standard error determination of the AUC which is computed by Var[TP]/n_1+Var[FP]/n_2 where Var[TP] is the variance of true positive results,$n_1$is the number of true positive results, Var[FP] is the variance of false positive results, and n_2 is the number of false positive results [37]. Despite apparently it has the advantage of the results to be centered on the absolute value, it was not recommended by Newcombe [38], claiming the feature of symmetry leads to an uncorrected calculus due to "overshoot" and "degeneracy." Newcombe suggested the Clopper-Pearson model for calculating exact binomial confidence intervals, used to the 95% CI determination, eliminating the incidence of both defects. The model contemplates the interval [AUC_{LL}, AUC_{HL}] with $AUC_{LL} \le p \le AUC_{HL}$, such that for all θ in the interval:

$$if\ AUC_{LL} \le \theta \le p, kp_r + \Sigma_{j:r<j\le n} p_{j\ge\alpha/2}; \text{and} \tag{16}$$

$$if\ p \le \theta \le AUC_{HL}, kp_r + \Sigma_{j:0\le j<n} p_j + kp_{r\ge\alpha/2} \tag{17}$$

where $j = 0, 1,\ldots, n$, R representing the random variable of which r is the realization, and $k = 1$ [39].

The significance level is equal to 5%, and it is another analogy to "diagnostic uncertainty" (see Section 2.1.4).

Since the AUC depends primarily on the number of true and false results, it is determined in RDT, considering the used "cutoff" exclusively.

2.2.3. Reproducibility, interferences, and re-evaluation

The reproducibility conditions, the interferences, and the re-evaluation are considered as seen in Sections 2.1.8, 2.1.9, and 2.1.10, respectively.

2.3. Agreement of binary results when comparator is other than diagnostic accuracy criteria

2.3.1. General concept

The principles used in Section 2.1 to compute the rates are not applicable to the determination of agreement between results of a candidate RDT and a comparative test. Despite the mathematical models being similar, the diagnostic accuracy is unmeasurable since the individuals' diagnostic is unidentified. It classifies merely the concordance of results. In this model, the diagnostic is replaced by the outcome of a comparative test different from a "gold standard" test [22]. The comparative test should have an acceptable diagnostic accuracy to minimize bias. When there is only available a comparative test without acknowledged diagnostic accuracy, the risk of bias effect in concordance is major, which could signify the agreements' results are unrealistic when compared to diagnostic accuracy. For example, this case happens if the comparative test's diagnostic accuracy is lower than that in the candidate RDT. In this condition, a nonconcordance could be erroneously interpreted. The evaluation is also erroneous when the agreement concept is misinterpreted as diagnostic accuracy.

Table 2 summarizes the agreement of binary results and the α and β errors.

Candidate test results	Comparative test		Total
	Positive ($x = 1$)	Negative ($x = 0$)	
Positive ($y = 1$)	Positive agreement (a)	Negative disagreement (b), α-error	$a + b$
Negative ($y = 0$)	Positive disagreement (c), β-error	Negative agreement (d)	$c + d$
Total	$a + c$	$b + d$	n

Table 2. 2 × 2 contingency table for test results agreement.

2.3.2. Agreement of overall results

The percentage of positive and negative results agreement is designated "overall percent agreement" *OPA* [%], and it is computed by (entry 10.2.1 of Ref. 22])

$$OPA[\%] = \frac{a+d}{n} \cdot 100 \tag{18}$$

where a is the total of candidate test's positive results among the positive results of the comparative test, d is the total of candidate test's negative outcomes among the negative results in the comparative test, and n is the total number of samples.

2.3.3. Agreement of positive results

The percentage of positive results agreement is designated "positive percent agreement" *PPA* [%], and it is computed by (entry 10.2.1 of Ref. [22])

$$PPA[\%] = \frac{a}{a+c} \cdot 100 \tag{19}$$

2.3.4. Agreement of negative results

The percentage of negative results agreement is designated "negative percent agreement" *NPA* [%], and it is computed by (entry 10.2.1 of Ref. [22])

$$NPA[\%] = \frac{d}{b+d} \cdot 100 \tag{20}$$

2.3.5. Inference of the results for the 95% confidence interval

Such as in the diagnostic accuracy (see Section 2.1.4) to infer an estimation to the population, a 95% confidence interval is used. Accordingly, the low and high limits for the test's overall percent agreement are computed respectively by (entry 10.2.2 of Ref. [22])

$$LL_{OPA}[\%] = \frac{Q_{1,OPA} - Q_{2,OPA}}{Q_{3,OPA}} \cdot 100 \tag{21}$$

$$HL_{OPA}[\%] = \frac{Q_{1,OPA} + Q_{2,OPA}}{Q_{3,OPA}} \cdot 100 \tag{22}$$

where $Q_{1,OPA} = 2 \cdot (a+d) + 1.96^2$, $Q_{2,OPA} = 1.96 \cdot \sqrt{1.96^2 + 4 \cdot (a+d) \cdot (b+c)/n}$ and $Q_{3,OPA} = 2 \cdot (n + 1.96^2)$.

Low and high limits for the test's positive percent agreement are computed respectively by (entry 10.2.2 of Ref. [22])

$$LL_{PPA}[\%] = \frac{Q_{1,PPA} - Q_{2,PPA}}{Q_{3,PPA}} \cdot 100 \tag{23}$$

$$HL_{PPA}[\%] = \frac{Q_{1,PPA} + Q_{2,PPA}}{Q_{3,PPA}} \cdot 100 \tag{24}$$

where $Q_{1,PPA} = 2 \cdot a + 1.96^2$, $Q_{2,PPA} = 1.96 \cdot \sqrt{1.96^2 + 4 \cdot a \cdot c/(a+c)}$, and $Q_{3,PPA} = 2 \cdot (a + c + 1.96^2)$.

Low and high limits for the test's negative percent agreement are computed, respectively, by (entry 10.2.2 of Ref. [22])

$$LL_{NPA}[\%] = \frac{Q_{1,NPA} - Q_{2,NPA}}{Q_{3,NPA}} \cdot 100 \tag{25}$$

$$HL_{NPA}[\%] = \frac{Q_{1,NPA} - Q_{2,NPA}}{Q_{3,NPA}} \cdot 100 \tag{26}$$

where $Q_{1,NPA} = 2 \cdot d + 1.96^2$, $Q_{2,NPA} = 1.96 \cdot \sqrt{1.96^2 + 4 \cdot b \cdot d / (b + d)}$, and $Q_{3,NPA} = 2 \cdot (b + d + 1.96^2)$ (note: *a*, *b*, *c*, and *d* are analogous to the number of "true positives," "false positives," "false negatives," and "true negatives" results of diagnostic accuracy).

The significance level is equal to 5%, and it is another time analogous to "diagnostic uncertainty." The claimed confidence interval should take into consideration the same limitations associated with the selection of intervals also discussed in Section 2.1.4.

2.3.6. Reproducibility and interferences

The reproducibility conditions and the interferences are treated as seen in Sections 2.1.8 and 2.1.9, respectively.

2.3.7. Re-evaluation

The RDT re-evaluation should take place mainly when significant epidemiological changes happen, and the comparative test is re-evaluated. As seen in Section 2.1.10, the variations in the batch of the reagent also involve re-evaluation study.

2.4. Seroconversion window period

2.4.1. General concept

The seroconversion window period is also recognized as the "seronegative period" or the "seroconversion sensitivity" [40]. Seroconversion period is defined as "the window period for a test designed to detect a specific disease (particularly an infectious disease) is the time between first infection and when the test can reliably detect that infection" [41]. The window period is equal to the number of days starting on the day of infection (day zero) to the day of the first positive result, that is, the time taken for seroconversion. Pereira et al. [42] proposed an alternative definition considering a trinary classification (i.e., positive/indeterminate/negative), which is uniquely applicable to tests with an ordinal scale of results.

2.4.2. Seronegative period

The evaluation of a seroconversion panel allows a primary determination of the most critical source of bias (biological bias), that is, the seronegative period. It is the major component of

the risk of the incorrect clinical decision [39]. The window period is also a diagnostic accuracy test despite not commonly classified as such to distinguish from Bayesian models (see Section 2.1).

The window period performance differs according to the type and the generation of the test. It depends also on the seroconversion panel used, which represents a unique infected individual and cannot be inferred to all the people with risk behaviors. Consequently, this period cannot be related to all the seronegative persons. It should not be misunderstood as the window period of a test, which is unknown. A benchmarking study should consider the shortest period in a rank of tests for comparison, usually coming with inserted literature in seroconversion panel.

For an in-depth discussion of seroconversion window period, further information can be found elsewhere [43].

2.4.3. Re-evaluation

The re-evaluation should only occur when a new panel with a shorter period is available. The laboratorian should consult the updated list of obtainable panels in the websites of manufacturers or the published literature.

3. Results and discussion

3.1. A case study

A practical example of the use of RDT results is in the evaluation of human blood components in a blood establishment (blood bank). In this case, the selection of the infected individuals should take into consideration the epidemiological prevalence not only of the tested agent but also types, subtypes, and other variants of significant epidemiological prevalence in the geographic area of the candidates to blood donors. If the sampling does not feature all the epidemiological significant variants, the diagnostic sensitivity is biased, and the results of the new variants are inaccurate. Otherwise, the selection of healthy individuals' sampling should include uniquely subjects recognized as noninfected. The regular blood donors' population is the best source for sampling of healthy subjects, since they are clinical and laboratory screened during a long term.

3.2. The rapid diagnostic test

To exemplify the models is used a single commercial RDT to detect antihuman immunodeficiency virus types 1 (HIV-1) and 2 (HIV-2) antibodies in human serum, plasma, venous blood, and capillary blood. Since it is to be used uniquely with serum, the evaluation occurs using serum specimens. The recombinant antigens (HIV-1 gp120, gp41, and HIV-2 gp36) conjugated to colloidal gold are adsorbed on a base of a strip of nitrocellulose membrane. There are

two test zones: (1) test zone, where HIV-1 and HIV-2 antigens are immobilized, and (2) control zone, where anti-HIV antibodies are immobilized. For internal quality control of the test is used a control zone. Interpret results as follows: there is no red strip in the test zone when the antibodies anti-HIV-1 or anti-HIV-2 are immeasurable, and there is a red strip in the control zone due to the migration of the antigen-conjugate gold (negative result). If the antibodies anti-HIV-1 or anti-HIV-2 are measurable, both zones have a red strip (positive result). Whenever the control zone does not have a red strip, the results do not have a binary classification (unclassified results). Repetitive positive must be tested using a confirmatory test, such as Western Blot, for antibodies.

3.3. Diagnostic accuracy I: 2 × 2 contingency tables

The evaluation contemplates a worst case considering:

- The laboratorians are trained successfully;
- The assessment is performed in a short period in reproducibility condition;
- The number of samples of infected individuals and healthy individuals is minor to assure a low cost of the assessment and to be feasible for a short period;
- The intended use of the RDT results is the screening of individuals with risk behavior to HIV infection.

Considering the major problem to a clinical decision to be false-negative outcomes, a larger number of infected individuals' samples is selected ($n = 10$), when compared to the number of healthy individuals ($n = 5$). The performance focuses on diagnostic sensitivity instead of diagnostic specificity. The samples are from diagnosed patients infected with HIV-1 ($n = 9$) and HIV-2 ($n = 1$), with repeated positive results to a fourth generation chemiluminescence immunoassay, Western Blot, and nucleic acid test (reverse transcription-polymerase chain reaction). Healthy subjects are regular blood donors tested in last three blood donations, with successive negative results to a fourth generation chemiluminescence immunoassay. The HIV group of the tested samples is not classified.

The claimed performance is determined according to the manufacturer's literature and the number of samples. When a 95% CI minimum requirement is claimed, the number of samples must be considered, since it is a restriction of the claim. Requested performance for absolute results is equal to 100% to diagnostic sensitivity and 90% to diagnostic specificity. The best diagnostic sensitivity's low limit is equal to 72.2%, and diagnostic specificity's low limit is equal to 56.6%, which is useless for the evaluation. In this case, the 95% CI is considered only to diagnostic sensitivity, since the minor number of healthy individuals' samples does not assure a reliable estimate. The lower limit of 95% CI of the claimed diagnostic sensitivity is 65%.

	Samples [*n*]		[%]							
	D_1	D_0	*Se*	*FNR*	*Sp*	*FPR*	*Eff*	*PPV*	*NPV*	*Pr*
Absolute	10	5	100	0	100	0	100	100	100	66.7
95% CI	–	–	72.2–100	–	56.6–100	–	–	–	–	–

Table 3. Number of samples and diagnostic accuracy of binary results.

Table 3 shows the performance rates to the candidate RDT, demonstrating that the requirements have been achieved with an overall accuracy of the test equal to 100%. **Figure 3** displays the true and false results in a three-dimensional data chart. The diameters of the disks are proportional to the number of outcomes. Despite the prevalence rates are unusable to the test evaluation, they could be useful to support the clinical decisions by the physician.

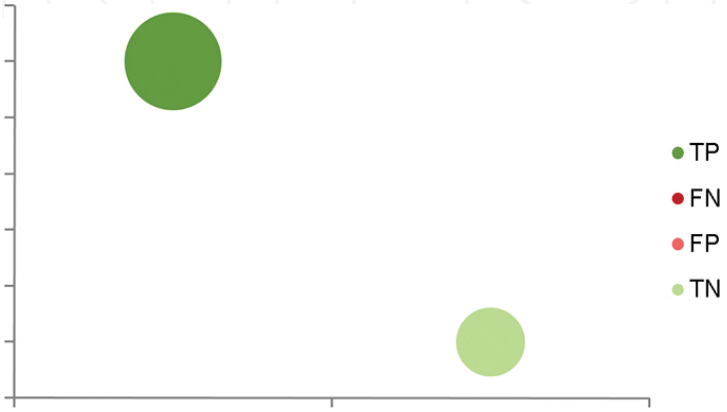

Figure 3. Bubble chart for the diagnostic uncertainty outcomes: true-positive (TP), false-negative (FN), false-positive (FP), and true-negative (TN) results.

It can be described that all samples of infected individuals and healthy individuals reported true results, and can be statistically considered that it is 95% confident about the inclusion of true-positive results in the interval or the error margin of 5% false-negative results attributed to that statement. The 95% CI can be interpreted as the diagnostic accuracy range.

The predictive value of positive and negative results is equal. This is interpreted as the samplings' results are accurately associated to the diagnostic.

3.4. Diagnostic accuracy II: receiver operating characteristic curve

The same data is used to determine the AUC as in the previous model. The AUC is computed using the Clopper-Pearson nonparametric method described above, since the hypothesis of the normal distribution of test results could not be tested. Claimed absolute value is 1.00 and the claimed 95% CI is [0.75, 1.00].

The AUC absolute value is 1.00, ranking the RDT as an "outstanding discrimination" test, with a 95% CI equal to [0.78, 1.00], whereby the RDT's AUC is accepted. It can be considered statistically that it is 95% confident about the inclusion of true results in the interval or the error margin of 5% false results attributed to that statement. Such as in Section 2.1.4., the 95% CI can be understood as the diagnostic accuracy interval. As it was already discussed, the AUC interpretation considers that the low limit is influenced by the number of samples tested. A minor number of samples decreases the statistical power of the evaluation, making its use unreliable.

These calculi are also useful for the definition of "cutoff" point in an RDT in the research and development phase. The "cutoff" is selected considering mainly its impact on the diagnostic sensitivity and specificity according to results' intended use, that could be interpreted as the adequation of the diagnostic sensitivity and specificity according to its role in the clinical decision. The measurements needed to plot the ROC curve are essential for manufacturers of *in vitro* diagnostics medical devices to determine the "cutoff" value for a test under development. This is also appropriate such as for medical laboratory "in house" and "nonwaived" modified tests (entry 42 CFR Subparts H, J, K and M of Ref. [44]), where the "cutoff" must be identified. The AUC single value is applicable only to the samplings of infected and healthy individuals, and it cannot be related to the population groups. For example, the "cutoff" point as a result of $AUC = 1.00$ divides perfectly the two groups, but it is unknown if it divides the two groups in the population. Then, the AUC 95% CI is determined since it is an inference to the population characterized in the same manner as the tested sampling. When the population has unlike characteristics than the sampling, the 95% CI is biased. The AUC utility, when applied to RDT evaluation in practice, is limited to represent the diagnostic test's accuracy in a single value, complementing diagnostic sensitivity and specificity results.

3.5. Agreement of binary results when comparator is other than diagnostic accuracy criteria

The same samplings as that in the previous models are used. They are also tested in a comparator test; in this example, a second RDT. Claimed requirements determine rates to be equal to 100% to the overall, positive, and negative agreements. The 95% score low confidence limit is claimed to be equal or higher than 70% to the overall and positive agreements. Such as in the 95% CI of diagnostic specificity, the confidence interval is not considered to the negative agreement. **Table 4** displays the estimates of results' agreement, showing that only the PPA result is achieved to the absolute and confidence interval. It can be statistically considered that it is 95% confident about the inclusion of agreements in the interval or the error margin of 5% disagreements attributed to that statement. In this case, the overall and negative agreements results were influenced by one false-negative result of the comparator test. This is an example of a major inaccuracy (bias) due to the use of a comparator with a weak performance (poor diagnostic specificity) when compared to candidate test. This evaluation approach should be only used exceptionally when the diagnosis is unspecified. The evaluator should use further studies to verify if the inaccurate results are due to the candidate's or comparator's test performance, to reduce the risk of erroneous conclusions.

	Samples [*n*]				
	Positive	Negative	*OPA* [%]	*PPA* [%]	*NPA* [%]
Absolute	10	5	93.3	100	83.3
95% CI	–	–	70.2–98.8	70.1–100	43.6–97.0

Table 4. Number of samples and agreement of binary results.

3.6. Seroconversion window period

The Window period estimate is performed with a seropositive commercial panel. The panel contains six undiluted and unpreserved plasma samples of an infected 19-year-old male. The manufacturer does not feature any RDT result in the inserted paper. **Figure 4** shows that 12 days after the first bleed, the RDT can detect the presence of HIV-1 antibodies reliably. Panel manufacturer inserted paper demonstrates that HIV RNA (copies/mL) is detected on the first day of bleed in one test in a series of two. Positive results are determined on the fifth day in three fourth generation immunoassays (detecting HIV-1 and HIV-2 antibodies, and HIV-1 p24 antigen) in a series of six tests. In a series of five antibody immunoassays, all have positive results after the 12th day. Thus, the RDT window period is equivalent to the tests to antibodies in this panel. As predictable, the window period was shorter in immunoassays also detecting HIV-1 p24 antigen and even shorter in nucleic acid tests. Such happens with any other test; a confirmatory scheme should be used to establish positive final results in RDT.

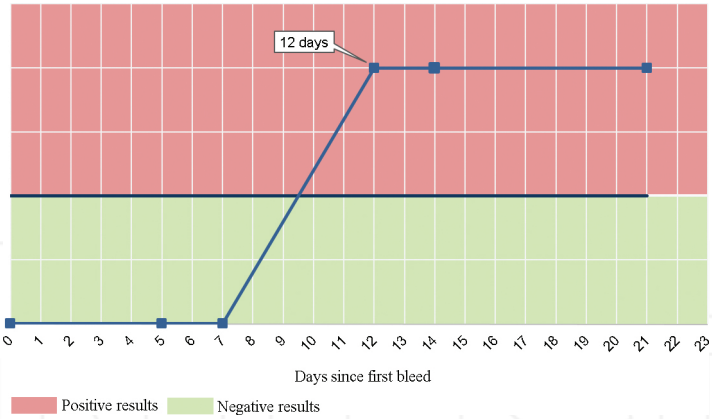

Figure 4. Results over time from six samples of a seroconversion commercial panel from a plasma donor who seroconverted over the course of their donation history.

3.7. Recommended flowchart

The selection and practical application of the evaluation models are summarized in a flowchart displayed in **Figure 5**.

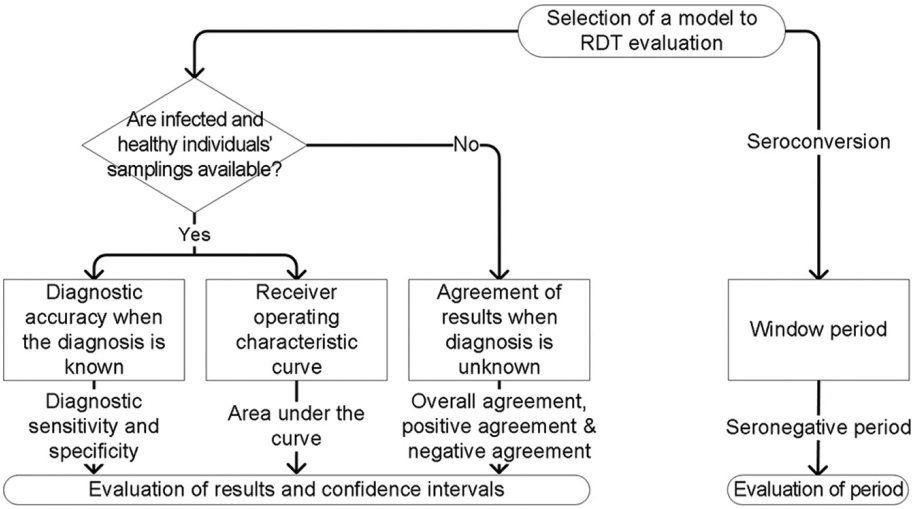

Figure 5. Flowchart to the selection of models to evaluate a RDT.

4. Conclusion

Summarizing, the diagnostic accuracy of RDT characterizes the level of confidence of binary results to be accurate (true) and the significance standards of a confidence interval. The determination of agreement of results is usually not a reliable estimate, for what its estimation should occur only when there is no possible to test samples with known diagnostic. The AUC is merely complementary to the diagnostic sensitivity and specificity evaluations. The seronegative window period is determined to evaluate an approximate period where false-negative results occur due to the antibodies' concentration being unmeasurable. This is expected in populations with a high epidemiological prevalence of tested agent. Such as in any other test, results have a chance to be untrue, whereby the physician should consider the predictive value of results during the clinical decisions based on the reported results.

Author details

Paulo Pereira

Address all correspondence to: paulo.pereira@ipst.min-saude.pt

Quality Management Department, Portuguese Institute of Blood and Transplantation, Lisbon, Portugal

References

[1] International Organization for Standardization. ISO 22870 Point-of-care testing (POCT)—Requirements for quality and competence. Geneva: ISO; 2006.

[2] Clinical and Laboratory Standards Institute. POCT04-A2 Point-of-care in vitro diagnostic (IVD) testing. 2nd ed. Wayne (PA): CLSI; 2006.

[3] Witonsky J. BioMarket Trends, IVD market moving rapidly on an upward trajectory. Genet Eng Biotechn 2015, 32(21):14–14. DOI: 10.1089/gen.32.21.05

[4] Reuters. Press release: Research and markets: Global point-of-care diagnostics market outlook 2014-2018—United States accounts for almost 50% of total market. Retrieved from: http://www.reuters.com/article/research-and-markets-idUSnBw206639a+100+BSW20140520. Accessed: March 26, 2016.

[5] Frost & Sullivan. NE3E-01 Analysis of the Global In Vitro Diagnostics Market. Retrieved from: http://www.frost.com/sublib/display-report.do?id=NE3E-01-00-00-00. Accessed: March 26, 2016.

[6] Bureau International des Poids et Mesures. JCGM 200:2012 International vocabulary of metrology—Basic and general concepts and associated terms, 2008 version with minor corrections. 3rd ed. Sèvres: BIPM; 2012. Retrieved from: http://www.bipm.org/utils/common/documents/jcgm/JCGM_200_2012.pdf. Accessed: March 26, 2016.

[7] International Organization for Standardization. ISO 31000 Risk management—Principles and guidelines. Geneva: ISO; 2009.

[8] Renaud B, Maison P, Ngako A, Cunin P, Santin A, et al. Impact of point-of-care testing in the emergency department evaluation and treatment of patients with suspected acute coronary syndromes. Acad Emerg Med 2008, 15(3):216–224. DOI: 10.1111/j.1553-2712.2008.00069.x

[9] Takakuwa K, Ou FS, Peterson E, Pollack, C, Jr, Peacock F, et al. The usage patterns of cardiac bedside markers employing point-of-care testing for troponin in non-ST-segment elevation acute coronary syndrome: Results from CRUSADE. Clin Cardiol 2009, 32(9):498–505. DOI: 10.1002/clc.20626

[10] Arias C, Panesso D, McGrath D, Qin X, Mojica M, et al. Rivaroxaban versus warfarin in nonvalvular atrial fibrillation. N Engl J Med 2011, 365(10):892–900. DOI: 10.1056/NEJMoa1009638

[11] Cohen D. Rivaroxaban: Can we trust the evidence? BMJ 2016, 352(i575):1–4. DOI: 10.1136/bmj.i575

[12] Van Hoof V. Quality in point-of-care testing. (How) can the quality of POCT results be guaranteed? In: Antwerp TQM conference; March 14 and 15, 2016; Belgium.

[13] International Organization for Standardization. ISO 9000 Quality management systems—Fundamentals and vocabulary. 4th ed. Geneva: ISO; 2015.

[14] Vesper H, Thienpont L. Traceability in laboratory medicine. Clin Chem 2009, 55(6): 1067–1075. DOI: 10.1373/clinchem.2008.107052

[15] Pereira P, Westgard J, Encarnação P, Seghatchian J, Sousa G. Quality management in the European screening laboratories in blood establishments: A view on current approaches and trends. Transfus Apher Sci 2015, 52(2):245–251. DOI: 10.1016/j.transci.2015.02.014

[16] Pereira P, Westgard J, Encarnação P, Seghatchian J, de Sousa G. The role of uncertainty in results of screening immunoassays in blood establishments. Transfus Apher Sci 2015, 52(2):252–255. DOI: 10.1016/j.transci.2015.02.015

[17] International Organization for Standardization. ISO 9001 Quality management systems—Requirements. 5th ed. Geneva: ISO; 2015.

[18] Clinical and Laboratory Standards Institute. EP29-A Expression of measurement uncertainty in laboratory medicine. Wayne (PA): CLSI; 2012.

[19] Clinical and Laboratory Standards Institute. H43-A2 Clinical flow cytometric analysis of neoplastic hematolymphoid cells. 2nd ed. Wayne (PA): CLSI; 2007.

[20] Elliott R, Jürgens R. Rapid HIV screening at the point of care: Legal and ethical questions. Canadian HIV/AIDS Legal Network; 2000. Retrieved from: http://www.aidslaw.ca/site/download/9359/. Accessed: March 26, 2016.

[21] Pepe M. The Statistical evaluation of medical tests for classification and prediction. Oxford: Oxford University Press; 2004.

[22] Clinical and Laboratory Standards Institute. EP12-A2 User protocol for evaluation of qualitative test performance. 2nd ed. Wayne (PA): CLSI; 2008.

[23] Zhou X, Obuchowski N, McCLish D. Statistical methods in diagnostic medicine. 2nd ed. Hoboken (NJ): John Wiley & Sons; 2002.

[24] Pereira P, Westgard J, Encarnação P, Seghatchian J. Evaluation of the measurement uncertainty in screening immunoassays in blood establishments: Computation of diagnostic accuracy models. Transfus Apher Sci 2015, 52(1):35–41. DOI: 10.1016/j.transci.2014.12.017

[25] Pereira P, Magnusson B, Theodorsson E, Westgard J, Encarnação P. Measurement uncertainty as a tool for evaluating the "grey-zone" to reduce the false negatives in immunochemical screening of blood donors for infectious diseases. Accred Qual Assur 2016, 21(1):25–32. DOI: 10.1007/s00769-015-1180-x

[26] Pereira P. Uncertainty of measurement in medical laboratories. In: Cocco L, editor. Measurement systems. Rijeka: InTech; 2016.

[27] Clinical and Laboratory Standards Institute. EP12-A User protocol for evaluation of qualitative test performance. Wayne (PA): CLSI; 2002.

[28] Young D. Effects of drugs on clinical laboratory tests. Volumes 1 & 2. 5th ed. Washington (DC): AACC Press; 2000.

[29] Young D. Effects of Disease on Clinical Laboratory Tests. Volumes 1 & 2. 4th ed. Washington (DC): AACC Press; 2001.

[30] Young D. Effects of Preanalytical Variables on Clinical Laboratory Tests. 3rd ed. Washington (DC): AACC Press; 2007.

[31] Narayanan S, Young D. Effects of herbs and natural products on clinical laboratory tests. Washington (DC): AACC Press; 2007.

[32] Clinical and Laboratory Standards Institute. EP24-A2 Assessment of the clinical accuracy of laboratory tests using receiver operating characteristic curves. 2nd ed. Wayne (PA): CLSI; 2011.

[33] D'Agostino RB, Belanger A, D'Agostino RB Jr. A suggestion for using powerful and informative tests of normality. Am Stat 1990, 44(4):316–321. DOI: 10.2307/2684359

[34] Mann H, Whitney D. On a test of whether one of two random variables is stochastically larger than the other. Ann Math Stat 1947, 18(1):50–60. DOI:10.1214/aoms/1177730491

[35] Metz C. Basic principles of ROC analysis. Semin Nucl Med 1978, 8(4):283–298. DOI: 10.1016/S0001-2998(78)80014-2

[36] Hosmer D, Jr, Lemeshow S, Sturdivant R. Assessing the fit of the model. In: Hosmer D Jr, Lemeshow S, Sturdivant R, editors. Solutions manual to accompany applied logistic regression. 2nd ed. Hoboken (NJ): John Wiley & Sons, 2013; p. 153–226.

[37] DeLong E, DeLong D, Clarke-Pearson D. Comparing the area under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988, 44(3):837–45. DOI: 10.2307/2531595

[38] Newcombe R. Two sided confidence intervals for the single proportion: A comparative evaluation of seven methods. Stat Med 1998, 17(8):857–872. DOI: 10.1002/(SI3CI)1097-0258(19980430)17:8<857::AID-SIM777>3.0.CO;2-E

[39] Clopper CJ, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika 1934, 26(4):404–413. DOI: 10.1093/biomet/26.4.404.

[40] Clinical and Laboratory Standards Institute. M53-A Criteria for laboratory testing and diagnosis of human immunodeficiency virus infection; approved guideline. Philadelphia (PA): CLSI; 2011.

[41] Le T, Bhushan V, Vasan N. First aid for the USMLE step 1. 20th ed. New York (NY): McGraw-Hill Medical; 2009.

[42] Pereira P, Westgard J, Encarnação P, Seghatchian J. Analytical model for calculating indeterminate results interval of screening tests, the effect on seroconversion window period: A brief evaluation of the impact of uncertain results on the blood establishment budget. Transfus Apher Sci 2014, 51(2):126–131. DOI:10.1016/j.transci.2014.10.004

[43] Fiebig E, Wright D, Rawal B, Garrett P, Schumacher R, et al. Dynamics of HIV viremia and antibody seroconversion in plasma donors: Implications for diagnosis and staging of primary HIV infection. AIDS 2003, 17(13):1871–1879. DOI: 10.1097/00002030-200309050-00005.

[44] Clinical Laboratory Improvement Amendments. CLIA law & regulations. Retrieved from: http://wwwn.cdc.gov/clia/Regulatory/default.aspx. Accessed: March 26, 2016.

INTECH

INTECH