The Journal of Molecular Diagnostics, Vol. 18, No. 6, November 2016

2016 JMD CME Program

ELSEVIER

the Journal of Molecular Diagnostics

jmd.amjpathol.org

REVIEW

Clinical Performance Evaluation of Molecular Diagnostic Tests

Bipasa Biswas

From the Center for Devices and Radiological Health, US Food and Drug Administration, Silver Spring, Maryland

CME Accreditation Statement: This activity ("JMD 2016 CME Program in Molecular Diagnostics") has been planned and implemented in accordance with the Essential Areas and policies of the Accreditation Council for Continuing Medical Education (ACCME) through the joint providership of the American Society for Clinical Pathology (ASCP) and the American Society for Investigative Pathology (ASIP). ASCP is accredited by the ACCME to provide continuing medical education for physicians.

The ASCP designates this journal-based CME activity ("JMD 2016 CME Program in Molecular Diagnostics") for a maximum of 36 *AMA PRA Category 1 Credit(s)*™. Physicians should claim only credit commensurate with the extent of their participation in the activity.

CME Disclosures: The authors of this article and the planning committee members and staff have no relevant financial relationships with commercial interests to disclose.

Accepted for publication June 1, 2016.

Address correspondence to Bipasa Biswas, Ph.D., Center for Devices and Radiological Health, US Food and Drug Administration, 10903 New Hampshire Avenue, Silver Spring, MD 20993. E-mail: bipasa.biswas@fda.hhs.gov.

Molecular diagnostic tests with application to clinical diagnostics involve studies in infectious diseases, inherited diseases, oncology, predisposition to disease, or the description of polymorphisms linked to disease states. General considerations in the design of evaluation of diagnostic test trials and statistical principles for reporting the results are discussed. A brief overview of the general statistical considerations related to the intent of use, test development versus validation, different types of biases, and issues with missing data are provided. Furthermore, issues related to commonly used but not necessarily correct methods to characterize the performance in the presence and absence of a clinical reference standard are discussed. These issues are broadly applicable to any molecular diagnostic test with a dichotomous result. This overview may help the clinical molecular diagnostic community to evaluate tests that provide a dichotomous result. *(J Mol Diagn 2016, 18: 803–812; http://dx.doi.org/10.1016/j.jmoldx.2016.06.008)*

Molecular diagnostic tests encompass a wide area of testing, such as testing for infectious diseases, oncologic tests, genetic tests for inherited diseases, and testing for predisposition to disease or polymorphisms linked to disease states, where the test involves detection of specific molecules, such as DNA, antibodies, or proteins. In the field of oncology, DNA tests have been used for screening for cancer (a multitarget stool DNA test for colorectal cancer screening[1]), microbial assays have been used to diagnose infectious diseases (assay for detection of group B *Streptococcus* in prenatal screening of specimens[2]), qualitative nucleic acid tests have been used for confirmation of hepatitis C virus infection and for screening blood donations,[3] and genetic tests have been used for inherited diseases (next-generation sequencing for cystic fibrosis transmembrane conductance regulator screening[4]). Molecular diagnostic test requires both analytical and clinical evaluations.[5–11]

Molecular diagnostics involve techniques to analyze biomarkers[12] in the genetic code of organisms, the genome, and how the cells express their genes as proteins, the proteomes.[5,12] These techniques apply molecular biology for medical testing to diagnose symptomatic individuals, screen asymptomatic individuals, monitor disease, provide prognosis in diseased patients, detect risk, and select patients for specific therapies. Molecular diagnostic tests use biological assays that detect a molecule, often in low concentrations, using PCR enzyme-linked immunosorbent assay or fluorescence *in situ* hybridization.[7–11,13,14] The detection of the biomarker uses real-time PCR, direct sequencing, or microarrays. Advances in next-generation sequencing will

Disclosures: None declared.

Portions of this work were presented at the 2006 US Food and Drug Administration/Industry Statistics Workshop held September 27–29, 2006, in Washington, DC.

Published by Elsevier Inc. on behalf of the American Society for Investigative Pathology and the Association for Molecular Pathology.
http://dx.doi.org/10.1016/j.jmoldx.2016.06.008

enable high-throughput DNA sequencing at relatively low cost for genomic-based diagnosis.[15]

Biomarker evaluation[12] by molecular diagnostics involves evaluation of both analytical performance and clinical performance. The analytical performance relates to the ability of the molecular diagnostic test to measure the underlying biological quantity under a variety of condition; although an important aspect of the test, it will not be discussed here. Several consensus standards are available to design and evaluate analytical performance of molecular diagnostic tests,[16–26] and useful resources are available from the US Food and Drug Administration's Center for Devices and Radiological Health Standards Program (US Food and Drug Administration, *http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/Standards/default.htm*, last accessed July 24, 2015). This review article focuses on the clinical performance evaluation for molecular diagnostic tests with dichotomous output. Clinical performance assesses the test's ability to detect the clinical or target condition of interest. A test result can be continuous, ordinal, or nominal.[27] A continuous or ordinal test result can be dichotomized[28] to give only two responses or output by using a cutoff. The test output for the molecular diagnostic test with dichotomous output is referred to as positive and negative in this review, which can also be interpreted qualitatively as the presence or absence of a target or clinical condition of interest. This article initially reviews general considerations, such as intent of use, development and validation, study conduct, and biases, then discusses possible performance measures and alludes to certain pitfalls of commonly used measures for performance evaluation, and finally discusses sample size justification and statistical analysis.

General Considerations in the Evaluation of Clinical Performance Trials

Clinical diagnostic performance for molecular diagnostic tests with dichotomous output is best evaluated with proper planning with respect to the intent of use, delineating development from validation, and adhering to appropriate study conduct to avoid potential sources of bias. Reporting of results is appropriately addressed by allowing for understanding of the study methods, the limitations involved, and correct interpretation of results.

Intent of Use

The intent of use of a molecular diagnostic test determines the type of study required to establish its performance. The intent of use describes the clinical purpose, the type of test, the criteria it measures, the specimen it measures (specimen type), the site of measurement, and the population for which the test is intended. Many variables can influence the performance of a test, such as population characteristics, the prevalence of the target condition of interest, the setting, and the type of test, among others. Thus, it is important to design the performance evaluation studies to match the intent of use. In general, it is important to include the following: the clinical purpose (eg, screening, diagnosis, prognosis, risk prediction, therapy or treatment selection for patients), target condition (eg, disease, disease stage, or any other condition of interest), target population, and the environment (eg, clinical laboratory, point of care, home use). Other important things to consider while designing a clinical study are anatomical location (eg, finger stick, venous) or specimen type from which the measurement is taken (eg, whole blood, plasma, serum, tissue), the measurand (which is being measured or detected), type of results (quantitative, continuous, ordinal, or qualitative) from the test, clinical interpretation of the test results, and the need for a trained or skilled user of the test and interpreter or reader of the test result.

Clinical Test Development and Validation

Medical tests often involve a number of technology and design parameters that are established in preclinical studies before conducting validation studies. For example, if the test is intended to be used qualitatively by dichotomizing the test result at a single cutoff or a decision threshold, then this has to be established before the final clinical validation study. This review article focuses on clinical performance of the molecular diagnostic test after finalization of all the design and technologic parameters, and thus considerations during the development are not the focus of further discussion.

A cutoff selection for a molecular diagnostic test with continuous or ordinal output may use the receiver operating characteristic (ROC) curve to select an optimum cutoff based on the clinical needs. The data set used to select an optimum cutoff is a training data set. An independent evaluation of the cutoff requires an assessment in an external data set that is independent and separate from that used in the selection of the cutoff. The ROC curve, for comparing two tests, provides additional support to discern whether a new test is better than a comparator test, although the test is to be used qualitatively by dichotomizing the test output. The ROC curve, which is a plot of 1—specificity and sensitivity on the *xy*-coordinate plane, helps to differentiate whether a new test is indeed on a different ROC curve that is superior to an existent test or whether the new test is just on the same ROC curve but that its operating point (cutoff or decision threshold) has been moved to provide a higher sensitivity at a loss of specificity. Further discussions related to cutoff selection at the development stage and the statistical techniques can be found in previously published articles.[25,26,28–30]

Once the test is finalized with regard to its design parameters and cutoff selection, the clinical performance is evaluated in a study population independent and separate from that used in the development of the test. Independent validation is desired because it objectively assesses the device performance external to the conditions and the data set used in development of the test and thus avoids issues related to training bias.

Study Conduct

Evaluation studies can be subject to many types of biases,[30−35] and careful consideration is needed at the study design stage and/or during analysis and reporting of performance to avoid potential sources of biases. Commonly observed sources of bias are selections bias, bias attributable to spectrum effect, verification bias, test evaluation bias, incorporation bias, imperfect reference test bias, and bias attributable to discrepant analysis. For example, tests can sometimes be evaluated using stored specimens collected from those with known target condition and can lead to inflated performance measures attributable to spectrum effect (when samples are collected from patients with a well-characterized target condition and healthy individuals without the target condition). Bias attributable to spectrum effect[34] occurs specifically when patients with the target condition of interest (eg, disease) in the study are not representative of the diseased patients in the intent-to-use population or conversely the nondiseased individuals are not from the intent-to-use population but rather healthy individuals. A common study design issue is to select patients with severe or chronic disease and patients who are healthier on average than nondiseased individuals in the population. Such designs can inflate the apparent accuracy of a diagnostic test.

Other sources of bias are selection bias in which individuals selected are not representative of those for whom the test will be applied. Bias attributable to spectrum effect is an example of selection bias. Another common example of selection bias occurs when a test is intended for screening the general population for a target condition but the study population instead consists of all individuals from referral site(s), who have been referred because of suspicion of the clinical or target condition of interest. The individuals in the study population from a referral site are not representative of the general screening population because the screening population would include many asymptomatic individuals. Biases attributable to sampling, such as selection bias or bias attributable to spectrum effect, are hard to quantify and cannot be addressed with a large sample size. Thus, sampling bias is best avoided or measures are taken to minimize bias at the study design stage.

Bias related to verification[30,35] occurs when a test's performance is restricted to individuals with definitive verification of the target or clinical condition of interest by the clinical reference standard. The magnitude of the bias is related to the association between selection for verification and the result of the test under evaluation. For example, if all individuals who test positive and only a few who test negative are verified by a clinical reference standard, then sensitivity will be biased upward, whereas specificity will be biased downward.

Bias can arise because of lack of masking, where the result of another test can influence the test procedure or its interpretation, which is very different from how the test will be applied in practice. This bias can be avoided by masking the test result from that of the comparator and the clinical reference standard and vice versa.

A test incorporation bias, which can be readily avoided, occurs when the result of the test is actually incorporated into the evidence used to diagnose the target or clinical condition. Because the evidence used for true diagnosis should be independent of the test under evaluation, such incorporations will bias the test performance.

A bias during analysis and reporting of performance can result from excluding patients for whom the diagnosis cannot be determined because of an intermediate, equivocal, or indeterminate test result. A planned analysis for reporting performance measures adequately addresses such biases.[36]

Misclassification by the clinical reference method introduces biases into the estimates of the performance measures, and one attempt to address these biases has been through discrepant analysis or discrepant resolution in which individuals with discordant test results by the index and comparator method are tested by a third resolving method, which may itself be imperfect or perfect. Variations on this design have been discussed, such as applying the resolver test only to patients with apparently false-negative results (those with positive results on the index test but negative results on the comparator test) or only to those with apparently false-positive results (those with the opposite discordancy). Although discrepant analysis of any form is intended to yield additional information about potentially problematic individuals, it introduces its own set of biases, always in an upward direction.[37,38]

Missing Data

Studies evaluating performance of molecular diagnostic tests could result in missing test results and/or the clinical reference standard. A missing clinical reference standard would result in verification bias, and reporting diagnostic performance that ignores missing clinical reference standard data can be misleading. The verification bias may occur in a predictable way if the decision to verify is based on observed test results or other clinical signs and symptoms but is in no way related to the underlying target or clinical condition being diagnosed. Thus, if the data selected for verification are based on a random sample of the observed test results and not the underlying target or clinical condition, the missing data can be imputed by multiplying the observed counts by inverse probability weighting (ie, inverse of the selection proportion). However, such procedures come with a penalty of generating less precise estimates of performance.

A missing test result could be because the result was invalid, the sample or specimen could not be obtained, or an informed consent could not be obtained before collecting the sample or specimen. Bias during analysis and reporting of performance can result from excluding patients with missing test results. Reporting the percentage of individuals with missing test results and the reasons the results are missing provides, at minimum, information on the performance of the test. In addition, planned analyses for reporting the test results and its performance measures addressing such biases[36] provide a comprehensive picture of the test's accuracy in the intent-to-use population.

When appropriate, imputation of missing data can be applied in statistical analyses to report performance measures.[39−41] Imputation is the procedure of filling up the missing data with plausible value, and several different imputation strategies are discussed in the article by Campbell et al.[39]

Performance Characteristics

The basic performance characteristics of a test are to inform how well the test measures what it intends to measure compared with a comparative benchmark (the clinical reference standard). For example, the basic performance measures to assess diagnostic accuracy of a qualitative test to distinguish diseased from nondiseased individuals involve sensitivity, that is, the probability that a truly diseased individual will test positive for disease, and specificity, that a truly nondiseased individual will test negative for disease. These measures are usually expressed as a percentage. Sensitivity and specificity are determined against a clinical reference standard that is used to identify individuals who truly have the clinical or target condition and those who do not have the clinical or target condition.

Errors in measuring the sensitivity and specificity of a test will arise if the reference standard itself is not accurate, that is, does not have 100% sensitivity and specificity, respectively, and is commonly known as an imperfect reference standard bias. Evaluating a diagnostic test is particularly challenging when there is no recognized clinical reference standard test. In the absence of a perfect reference standard, performance of a test evaluated against an imperfect reference standard is expressed as positive percent agreement (PPA) (the proportion of individuals with the target condition by the imperfect reference standard who test positive) and negative percent agreement (NPA) (the proportion of individuals free of the target condition by imperfect reference standard who test negative). Both PPA and NPA are reported as percentages.

Overall percentage agreement is often reported in lieu of sensitivity-specificity pair or PPA-NPA pair. Overall agreement is not independent of the prevalence of the target condition; thus, for a low prevalence, the overall agreement may look good, although the test performs poorly for detecting the target condition. Thus, overall agreement is not acceptable to evaluate a test performance.[42]

Two other important measures of test performance (when evaluated against a perfect reference standard) are positive predictive value (PPV), the probability that those testing positive by the test truly have the disease, and the negative predictive value (NPV), the probability that those testing negative by the test are truly nondiseased. Both PPV and NPV depend not only on the sensitivity and specificity of the test but also on the prevalence of disease in the population studied.

Clinical Performance Evaluation

A clinical performance study[43−45] of a diagnostic test provides information about diagnostic accuracy in a clinical setting and on a study population for whom the test is intended. The design of a study is greatly improved if the process is approached systematically by defining the need and the objectives of an evaluation trial[46]; defining the type of trial, study population, site selections, and conduct of trial; and finally reporting the results.[47,48]

Trial Objectives

The purpose of the study and the trial objective in the clinical context is usually defined well in advance of the conduct of a clinical performance evaluation study. Performance evaluation of the diagnostic test usually includes a prespecified performance goal in terms of diagnostic accuracy. The goals could be based on prespecified minimal targets for sensitivity and specificity for evaluating a single molecular diagnostic test weighing the benefits of a correct test result and the risk associated with false results (false positive and false negative) in the clinical context. The intent of the trial could be to replace an old test with a new test to perform a direct comparison based on sensitivity and specificity. On the other hand, if the intent of a trial is to compare a new test with another test, in the absence of an established clinical reference standard, such a study could involve performance based on minimal targets for PPA and NPA.

Design of the Diagnostic Trial

The design of a clinical performance evaluation trial for a molecular diagnostic test depends on the purpose of the trial and the population for which the test is intended. Defining the target population is best served by taking into account the probable purpose of the test. For example, will it replace an existing test, triage patients in need of further investigation, or be used as an additional test in a diagnostic test strategy?

A prospectively designed cohort enrolling a consecutive or random sample from the target population for whom the molecular diagnostic test will be used in actual clinical practice is desirable for a study population. Individuals who meet the inclusion and exclusion criteria are selected consecutively or a random sample from the target population is selected. A well-designed and well-executed prospective study can ensure that the study population provides an adequate representation of the target population and often provides the highest-quality evidence of the test performance. Often, for rare target conditions, such prospectively planned studies have to be large to get a sufficient number of patients with the target condition of interest. However, retrospective convenience sampling using stored specimens collected from those with available samples and with reference standard results can introduce selection bias. Tests evaluated on stored specimens collected from those with and without the known target condition can lead to inflated test performance. The spectrum effect can occur when the stored samples are from sickest of the sick and healthiest of the healthy. Individuals between these two extremes usually have conditions that are more difficult to diagnose and are excluded from such studies, leading to inflated test performance.

Although a prospectively planned clinical study provides the highest quality of evidence, it may not always be feasible. Alternatively, a prospectively planned analysis of retrospective archived specimens from a well-conducted and recently completed study can provide an efficient and convenient alternative.[49] In a prospective-retrospective study design, the test result is obtained on archived material and then examined for a prespecified trial objective. Study design includes statistical justification for adequacy of sample size based on the study objectives and plans on how to handle, prepare, process, and select archived material. It is essential that the prospective planners of the analysis be blinded to the data from archived material, and that the blinding be documented in the study protocol. However, before planning such a study, an important step is to characterize the analytical performance of the test before it is applied to the archived material. Otherwise, the archived material will be wasted on a test with poor analytical performance. The test results on archived material may not be completely concordant with those from fresh material. The archived material may not be available for all participants, the archived material may deteriorate over time, or the test may not be able to provide a result on some archived material. Retrospective sampling of individuals with available specimen and reference standard results can introduce bias attributable to the spectrum effect. Thus, the level of evidence from a prospective-retrospective study is typically not considered to be of high enough quality in regard to test performance compared with a prospective study. In addition, selection of participants and testing of specimens need to be conducted in such a way that the performance of the test compared with the clinical reference standard is not confounded by analytical variables, such as day, user, reagent lot, collection site, and testing site, or other ancillary variables that may be associated with the test and clinical reference standard. Nonetheless, well-designed and carefully planned prospective-retrospective study with statistical rationale for sample size and prespecified analysis plan and plans for handling, preparing, and processing archived material is prone to fewer sources of bias than convenience sampling.

A well-designed trial usually specifies the setting where participants are enrolled and where the test will be conducted. The setting could be in a specialized clinic or laboratory, a point-of care-setting (eg, hospital), or at home. Tests will probably be performed differently, depending on the setting.

Performance evaluation of tests is based on a comparator method, which is usually a clinical reference standard. The choice of an appropriate reference standard is crucial for the evaluation of a test for diagnostic accuracy. The evaluation usually involves a paired design by which each participant is evaluated by the test and also the clinical reference standard. In the absence of a reference standard, the performance of a test against an imperfect comparator is reported using the PPA and NPA. However, agreement measures are not the same as sensitivity and specificity of a test, and a pitfall of such measures to evaluate agreement is that a perfect test when compared against an imperfect reference standard may indicate less than perfect agreement with the imperfect reference standard.[50]

When diagnosing the presence or absence of a target condition, the performance of the test can vary from study to study; thus, the performance measure pairs (sensitivity-specificity, PPA-NPA, PPV-NPV) are evaluated in the same study. Likewise, while comparing the performance of two tests, comparing two tests evaluated in two separate studies is misleading because the study populations are not the same; thus, there could be potential imbalances in the spectrum of participants between the two separate studies.

When comparing the performance of two tests, the comparative study designs commonly involve a parallel group design or a paired design among other designs. In a parallel group design, participants are randomized to two groups for evaluation by one of the two tests but not both, and each participant is also verified for the target condition by the reference standard. Parallel group designs may be useful when molecular diagnostic tests are invasive, and it may not be feasible to apply more than one test on an individual. Even when randomization is used, it is possible to have imbalance across groups in terms of spectrum of the target condition being evaluated. A parallel group design essentially follows the same design principles as that of a parallel arm randomized clinical trial so that the participants are randomized to two different tests and the randomization should ensure that the study arms are balanced with regard to factors that affect the test performance. Sample sizes for such designs would have to be much larger than for paired designs to overcome potential imbalance attributable to variability in target condition being evaluated. For comparison that involves a paired design, also referred to as three-way comparison, individuals receive both tests and also the reference standard. Comparisons from the paired design are usually more statistically efficient than those from an equivalent parallel group design because variability across participant groups evaluated for by the two tests introduces additional imprecision to the parallel group design. An advantage of such a design is that possibilities of confounding are eliminated. In addition, one can examine the particular characteristics of individuals with two different test results.[31]

Performance Measures

A clinical performance evaluation usually includes the outcomes of the evaluation, such as performance measures with prespecified performance goals or a comparison of a new test to an old test based on prespecified effect sizes. A diagnostic accuracy as performance measure provides well-characterized information of the performance of a test when evaluated against a reference standard.

Performance evaluation of a single test against a clinical reference standard would usually require identifying a minimally acceptable sensitivity and specificity to design such studies.

A study in which a reference standard requires invasive techniques or a screening study on a large study population with large number of participants with negative test results may lead to fewer individuals verified by a reference standard, particularly for those with negative results. Performance measures, such as the sensitivity-specificity pair, based on only individuals with an available reference standard are misleading because these measures are biased. Adequate planning is required to address limited verification by reference standard of individuals in the study population at the study design stage. If verification of the target condition by the reference standard is in no way related to the true target condition other than the observed test result, the information missing can be considered as missing at random. Unbiased estimates of clinical performance of the test can be obtained by appropriate imputation or inverse probability weighting.[30,31,51] In addition, a comparison of ratios of sensitivity and ratios of 1—specificity may provide information on whether the replacement test is better than an existing test.[51–53]

If reference standard is available and all participants can be verified for determining the true target condition, a comparison of performances is based on sensitivity of the two tests and likewise for specificity of the two tests.[30,31]

Inappropriate Statistics or Tests to Evaluate Clinical Performance of Diagnostic Tests

Often overall agreement and/or κ statistics are used to evaluate agreement between two tests. Although overall agreement provides an agreement between two tests, it fails to differentiate the agreement of positive results with the presence of the target or clinical condition and agreement of negative results with the absence of the target or clinical condition.

An overall agreement as an evaluation of performance of a diagnostic test is influenced by the prevalence of the target condition.[42] Thus, mathematically, if T denotes the test and R denotes the clinical reference standard, then $T+$ $(R+)$ and $T-$ $(R-)$ denotes test T (Reference R) positive and test (reference) negative. The influence of prevalence on overall agreement can be further illustrated. If p $(= Pr\ (R+))$ denotes the prevalence and

$$\pi_{se}\left[=Pr(T+\ |R+)\right] \text{ and} \tag{1}$$

$$\pi_{sp}\left[=Pr(T-\ |R-)\right] \tag{2}$$

denote the sensitivity and specificity of a test, respectively, then the overall percent agreement is as follows:

$$\begin{aligned} OA = Pr\,(T=R) &= Pr\,(T+\ |R+)\,Pr\,(R+) \\ &+ Pr\,(T-\ |R-)\,Pr\,(R-) = p * \pi_{se} + (1-p) * \pi_{sp} \end{aligned} \tag{3}$$

If p, the prevalence of the target condition, is very small, a test with low sensitivity but with a high specificity will result in a high overall agreement, although the test is not good at detecting the target condition.

Similarly, κ statistics to evaluate agreement are sensitive to the prevalence of the target condition.[42,54,55] The κ statistics are mathematically defined as follows:

$$\begin{aligned} \kappa &= \frac{\Pr(T=R)-[\Pr(T+)\Pr(R+)+\Pr(T-)\Pr(R-)]}{1-[\Pr(T+)\Pr(R+)+\Pr(T-)\Pr(R-)]} \\ &= \frac{2p(1-p)\left[\pi_{se}+\pi_{sp}-1\right]}{1-\left[p^2\pi_{se}+(1-p)^2\pi_{sp}+p(1-p)\left(2-\pi_{se}-\pi_{sp}\right)\right]} \end{aligned} \tag{4}$$

A test with the same sensitivity and specificity (ie, $\pi_{se} = \pi_{sp}$) will yield the maximum κ at $P = 0.5$, and the κ decreases if $P < 0.5$ or $P > 0.5$. A test with lower sensitivity than specificity (ie, $\pi_{se} < \pi_{sp}$) will yield a higher κ than when the sensitivity and specificity are switched (ie, $\pi_{se} > \pi_{sp}$) in the same study population with prevalence $P < 0.5$. Similarly, a test with higher sensitivity than specificity (ie, $\pi_{se} > \pi_{sp}$) will yield a higher κ on a study population with prevalence $P > 0.5$ than a test for which sensitivity and specificity are switched (ie, $\pi_{se} < \pi_{sp}$). Thus, κ is not independent of the prevalence of the target condition.

The McNemar χ^2 test[56] for paired study design does not appropriately[42] assess agreement between a test and an imperfect comparator. The McNemar χ^2 test assumes a null hypothesis that the rates of positive responses by the two tests are equal.[56] The McNemar χ^2 test could lead to the conclusion that there is not enough evidence to demonstrate that the two medical tests differ, when in truth the two differ. Alternatively, the two medical tests may have very high agreements, and yet the McNemar test rejects that the two are equal.[42] In summary, overall agreement and κ statistics do not appropriately measure agreement between two medical tests and are inappropriate as primary measures of evaluation for agreement. Likewise, the McNemar χ^2 test to evaluate agreement is also not recommended.[42]

Sample Size

The key question to address before any clinical study is what level of performance is required of a test. Increasing the sample size reduces the uncertainty regarding the performance measures where the extent of uncertainty is summarized by the CIs.

For evaluating a single test against a reference standard, the sample size for a clinical study is based on performance goals. Sample size affects the width of the CI, and the narrower the width, the greater the precision of the estimate. Sample size also affects the statistical power (probability of rejecting a false null hypothesis) associated with hypothesis test for test performance. In addition, sample size for a prospectively designed cohort accommodates for the prevalence of the target condition.[30,31]

For evaluating the performance of molecular diagnostic test based on a minimally acceptable sensitivity prespecified by se_0 and a minimally acceptable specificity prespecified by sp_0, respectively, the statistical hypotheses statements for sensitivity and specificity are as follows:

$$\begin{aligned} &H0\ (\textit{null hypothesis}){:}\ se \leq se_0 \\ &H1\ (\textit{alternative hypothesis}){:}\ \ se > se_0 \end{aligned} \tag{5}$$

$$\begin{aligned} &H0\ (\textit{null hypothesis}){:}\ sp \leq sp_0 \\ &H1\ (\textit{alternative hypothesis}){:}\ \ sp > sp_0 \end{aligned} \tag{6}$$

The goal of selecting the sample size is to ensure that the molecular diagnostic test meets acceptable performance and the study has high probability (power) to detect that the tests are better than the minimal acceptable performances. Sample size calculation based on prespecified minimally acceptable sensitivity and specificity would thus require a value of sensitivity (se_1) and specificity (sp_1) under the alternative hypothesis. A general formula for computing the sample size for a test that tests these hypotheses is as follows:

$$n_D = \frac{\left[z_{1-\alpha}\sqrt{se_0(1-se_0)} + z_{1-\beta}\sqrt{se_1(1-se_1)}\right]^2}{[se_1 - se_0]^2} \tag{7}$$

$$n_{ND} = \frac{\left[z_{1-\alpha}\sqrt{sp_0(1-sp_0)} + z_{1-\beta}\sqrt{sp_1(1-sp_1)}\right]^2}{[sp_1 - sp_0]^2} \tag{8}$$

where n_D is the number of individuals with the clinical or target condition, n_{ND} is the number of individuals without the clinical or target condition, $z_{1-\alpha}$ is the $1-\alpha$ percentile of a standard normal distribution, α is the type I error rate, $z_{1-\beta}$ is the $1-\beta$ percentile of a standard normal distribution, and β is the type II error rate (or 1-power).

For example, if an investigator sets a minimally acceptable sensitivity at 70.0% and $\alpha = 0.025$ ($z_{1-\alpha} = z_{0.975} = 1.96$) and power is 80.0% ($\beta = 0.20$, $z_{1-\beta} = z_{0.80} = 0.84$) to detect a sensitivity of 80.0%, then the sample size required for individuals with the clinical or target condition (disease) is as follows:

$$\begin{aligned} n_D &= \frac{\left[1.96\sqrt{0.70(1-0.70)} + 0.84\sqrt{0.80(1-0.80)}\right]^2}{[0.80-0.70]^2} \\ &= \frac{\left[1.96\sqrt{0.70(0.30)} + 0.84\sqrt{0.80(0.20)}\right]^2}{[0.10]^2} \approx 153 \end{aligned} \tag{9}$$

When comparing two tests, sample size is based on the effect size of the test performances.[30,31,57–60] The comparisons could be to evaluate superiority, equivalence, or noninferiority between a new test and a comparator test, depending on the clinical need that the new test satisfies. If se_T and se_C denotes the sensitivity of a new test and the comparator test, then the null and alternative hypothesis statements for superiority are as follows:

$$\begin{aligned} &H0\ (\textit{null hypothesis}){:}\ se_T \leq se_C \\ &H1\ (\textit{alternative hypothesis}){:}\ \ se_T > se_C \end{aligned} \tag{10}$$

Similarly, hypothesis can be stated for specificity.

The goal of selecting the sample size must be to ensure that if the new test (T) truly is superior to the comparator test (C), then the study will have a high probability (power) to detect the difference. A sample size calculation for comparing diagnostic accuracy of two molecular diagnostic tests would require a value of the difference between sensitivities ($se_T - se_C$) denoted by, for example, Δ under the alternative hypothesis. A general formula for computing the sample size for number of patients with target condition for a test that tests these hypotheses is as follows:

$$n_D = \frac{\left[z_{1-\alpha}\sqrt{V_0\left(\widehat{se_T} - \widehat{se_C}\right)} + z_{1-\beta}\sqrt{V_1\left(\widehat{se_T} - \widehat{se_C}\right)}\right]^2}{[\Delta]^2} \tag{11}$$

where $z_{1-\alpha}$ is the $1-\alpha$ percentile of a standard normal distribution, α is the type I error rate, $z_{1-\beta}$ is the $1-\beta$ percentile of a standard normal distribution, β is the type II error rate (1-power), $V_0(\widehat{se_T} - \widehat{se_C})$ is the variance function of the estimated difference is sensitivity of the two tests under the null hypothesis, and $V_1(\widehat{se_T} - \widehat{se_C})$ is the variance function of the estimated difference or the sensitivity of the two tests under the alternative hypothesis.

The variance functions for a three-way comparison in a paired design (when participants undergo both tests and are verified by a reference standard test) takes the following general form:

$$\begin{aligned} V\left(\widehat{se_T} - \widehat{se_C}\right) &= V\left(\widehat{se_T}\right) + V\left(\widehat{se_C}\right) - 2Cov\left(\widehat{se_T}, \widehat{se_C}\right) \\ &= se_T \times (1 - se_T) + se_C \times (1 - se_C) - 2Cov\left(\widehat{se_T}, \widehat{se_C}\right) \end{aligned} \tag{12}$$

where $Cov(\widehat{se_T}, \widehat{se_C})$ is the covariance function of $\widehat{se_T}$ and $\widehat{se_C}$ equaling zero for studies in which different patients are evaluated by the two tests (ie, parallel group design). Thus, for a parallel group design, n_D is the sample size of the number of patients with the target condition in one group, and the total number of patients with the target condition in the parallel group study is $2 \times n_D$.

For a three-way comparison, matched-pair study in which the participants undergo evaluations by both the new test and the comparator test, the variance function is given as follows[56]:

$$V_0\left(\widehat{se_T}-\widehat{se_C}\right)=\psi \tag{13}$$

$$V_1\left(\widehat{se_T}-\widehat{se_C}\right)=\psi-\Delta^2 \tag{14}$$

where

$$\psi=se_T+se_C-2\times se_C\times P(T=1|C=1) \tag{15}$$

where $P\ (T = 1|C = 1)$ is the probability that the result of the new test (T) is positive given that the result of the comparator test (C) is positive and se_T and se_C are the conjectured values of sensitivity from the alternative hypothesis. The value of ψ ranges from Δ (when the correlation between two test results is perfect) to $se_T \times (1 - se_C) + se_C \times (1-se_T)$ (when the two test results are independent). Without any information about the value of the correlation between the two tests, it is prudent to use a sample size that ensures adequate power. Similar calculations can be performed for specificity to determine the sample size for the number of individuals without the clinical or target condition to adequately power the study.

A prespecified approach to the performance goal, study design, statistical hypotheses, and statistical analysis plan lends to the credibility of the study. Choosing a statistical analysis plan to accommodate the data inflates the chance of making a false conclusion and is strongly discouraged. Any readjustment of sample size after unmasking the data without prior specification and adequate adjustments of α-spending for multiple looks is strongly discouraged because it violates the basic principles of hypothesis testing. A prespecified analysis plan with appropriate adjustments of α-spending for any interim analysis adheres to the highest statistical rigor.

Reporting

Complete and accurate reporting of the performance measures and CIs to characterize the uncertainty of the estimates provides necessary information for evaluation of medical test outcomes.[43,47] The sensitivity-specificity pair (when the test is evaluated against a clinical reference standard) or the PPA-NPA pair (when the test is evaluated against an imperfect comparator that is not a clinical reference standard) is reported along with respective CIs.[61]

The usual approach to constructing a CI for a measure of diagnostic accuracy assumes a large sample size, so it is reasonable for the measure to follow a normal (or Gaussian) distribution. Thus, an asymptotic 100% $(1 - \alpha)$ CI for sensitivity is as follows:

$$\widehat{Se}-z_{1-\alpha/2}\sqrt{\widehat{Var}\left(\widehat{Se}\right)},\quad \widehat{Se}+z_{1-\alpha/2}\sqrt{\widehat{Var}\left(\widehat{Se}\right)} \tag{16}$$

The CI for specificity is formed similarly.

However, the above formula has major drawbacks. The percentage of time that the CI actually includes the true value of the accuracy measure is much smaller than desired, particularly for small sample size or for accuracy measures close to 0 or 1. Alternatively, exact CIs for sensitivity and specificity can be computed from the binomial distribution. Alternatively, a score CI[62] can be reported. The confidence limits for sensitivity are as follows:

$$\frac{\widehat{Se}+z^2_{1-\alpha/2}\Big/(2n_D)\pm z_{1-\alpha/2}\sqrt{\left[\widehat{Se}\left(1-\widehat{Se}\right)+z^2_{1-\alpha/2}\Big/4n_D\right]\Big/n_D}}{1+z^2_{1-\alpha/2}\Big/n_D} \tag{17}$$

where $\widehat{Se}$ is the sensitivity, estimated from the study with n_D patients with the clinical or target condition and $z_{1\ -\ \alpha/2}$ is the $1 - \alpha/2$ percentile of a standard normal distribution. The confidence limits for specificity can be calculated as follows:

$$\frac{\widehat{Sp}+z^2_{1-\alpha/2}\Big/(2n_{ND})\pm z_{1-\alpha/2}\sqrt{\left[\widehat{Sp}\left(1-\widehat{Sp}\right)+z^2_{1-\alpha/2}\Big/4n_{ND}\right]\Big/n_{ND}}}{1+z^2_{1-\alpha/2}\Big/n_{ND}} \tag{18}$$

where $\widehat{Sp}$ is the specificity, estimated from the study with n_{ND} individuals without the clinical/target condition, and $z_{1\ -\ \alpha/2}$ is the $1 - \alpha/2$ percentile of a standard normal distribution.

An example from literature[63] describes a rapid recombination polymerase amplification assay for detection of *Chlamydia trachomatis* in human urine samples. A sensitivity of 83.3% in 12 patients with *C. trachomatis* positive urine samples was reported; thus, on the basis of the above formula with $\alpha = 0.05$ ($z_{1\ -\ \alpha/2} = z_{0.975} = 1.96$) and $n_D = 12$, a CI of 95% for sensitivity of the assay was 55.2%−95.3%. If instead the study had enrolled 200 participants with *C. trachomatis* positive urine samples, a CI of 95% would be 77.5%−87.8% for same sensitivity (83.3%) of the assay. A specificity of 100.0% in 58 individuals with *C. trachomatis* negative urine samples was reported; thus, on the basis of the above formula with $\alpha = 0.05$ ($z_{1\ -\ \alpha/2} = z_{0.975} = 1.96$) and $n_{ND} = 58$, a CI of 95% for specificity of the assay was 93.8%−100.0%. If instead the study had enrolled 200 individuals with *C. trachomatis* negative urine samples, a CI of 95% would be 98.1%−100.0% for same specificity (100.0%) of the assay. Thus, with increased sample size, the uncertainty in the estimates decreases.

The PPV and NPV[30,31,57,64] can be used in prospective evaluation but with the caution that these depend on the prevalence of the target condition being evaluated and the test has been evaluated against a clinical reference standard.

Generally, it is recommended that PPV-NPV be provided for a set of prevalence values.

It is necessary to understand the study findings, limitations if any for generalizability to the target population, potential sources of bias, and analyses accounting for missing data. Many recommendations are available for appropriate reporting of evaluation studies for medical tests.[43,47]

Conclusion

This review provides an overview of general considerations for clinical evaluation of molecular diagnostic tests with dichotomous output, discussing study designs and general statistical considerations for test performance. Clinical performance evaluation studies could be designed to evaluate diagnostic accuracy or as a replacement for existing tests. Careful considerations for intent of use of the test, study population, trial objective, and reference standard are among many required to ensure validity of the study outcomes, and design should avoid potential sources of bias.

Careful consideration and planning are required for accurate reporting of the study outcome. Diagnostic accuracy (sensitivity-specificity), when a test is evaluated against a reference standard, and agreement measures (PPA and NPA) in the absence of a reference standard are recommended for performance evaluation.

Acknowledgment

I thank Lakshmi R. Vishnuvajjala, Ph.D., for her critical review of the manuscript.

References

1. Imperiale TF, Ransohoff DF, Itzkowitz SH, Levin TR, Lavin P, Lidgard GP, Ahlquist DA, Berger BM: Multitarget stool DNA testing for colorectal-cancer screening. N Engl J Med 2014, 370:1287−1297
2. Buchan BW, Faron ML, Fuller D, Davis DE, Mayne D, Ledeboer NA: Multicenter clinical evaluation of the Xpert GBS LB assay for detection of group B Streptococcus in prenatal screening specimens. J Clin Microbiol 2015, 53:443−448
3. Scott JD, Gretch DR: Molecular diagnostics of hepatitis c virus infection: a systematic review. JAMA 2007, 297:724−732
4. Loukas YL, Thodi G, Molou E, Georgiou V, Dotsikas Y, Schulpis KH: Clinical diagnostic next-generation sequencing: the case of CFTR carrier screening. Scand J Clin Lab Invest 2015, 75: 374−381
5. Grody WW, Nakamura RM, Strom CM, Kiechle FL: Molecular Diagnostics: Techniques and Applications for the Clinical Laboratory. Boston, MA, Academic Press Inc, 2010
6. Carl AB, Edward RA, David EB: Tietz Textbook of Clinical Chemistry and Molecular Diagnostics. Amsterdam, the Netherlands: Elsevier, 2012
7. Establishing Molecular Testing in Clinical Laboratory Environments; Approved Guideline. CLSI document MM19-A. Wayne, PA: Clinical and Laboratory Standards Institute, 2011
8. Nucleic Acid Sequencing Methods in Diagnostic Laboratory Medicine; Approved Guideline—Second Edition. CLSI document MM09−A2. Wayne, PA: Clinical and Laboratory Standards Institute, 2014
9. Nucleic Acid Amplification Assays for Molecular Hematopathology; Approved Guideline—Second Edition. CLSI document MM05−A2. Wayne, PA: Clinical and Laboratory Standards Institute, 2012
10. Molecular Methods for Clinical Genetics and Oncology Testing; Approved Guideline—Third Edition. CLSI document MM01-A3. Wayne, PA: Clinical and Laboratory Standards Institute, 2012
11. Quantitative Molecular Methods for Infectious Diseases; Approved Guideline—Second Edition. CLSI document MM06-A2. Wayne, PA: Clinical and Laboratory Standards Institute, 2010
12. Biomarkers Definitions Working Group: Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. Clin Pharmacol Ther 2001, 69:89−95
13. Rao JR, Fleming CC, Moore JE (Eds): Molecular Diagnostics: Current Technology and Applications. Nofolk, UK: Horizon Bioscience, 2006
14. Van Ommen GJB, Breuning MH, Raap AK: FISH in genome research and molecular diagnostics. Curr Opin Genet Dev 1995, 5: 304−308
15. Voelkerding KV, Dames SA, Dusrstchi JD: Next-generation sequencing: from basic research to diagnostics. Clin Chem 2009, 55: 641−658
16. Accuracy (trueness and precision) of Measurement Methods and results—Part 1: General Principles and Definitions, ISO 5725−1. Geneva: International Organization for Standardization, 1994
17. Accuracy (trueness and precision) of Measurement Methods and Results—Part 2: Basic Method for the Determination of Repeatability and Reproducibility of a Standard Measurement Method, ISO 5725−2. Geneva: International Organization for Standardization, 1994
18. Measurement Procedure Comparison and Bias Estimation Using Patient Samples; Approved Guideline—Third Edition. CLSI document EP09−A3. Wayne, PA: Clinical and Laboratory Standards Institute, 2013
19. Evaluation of Precision of Quantitative Measurement Procedures; Approved Guideline—Third Edition. CLSI document EP05-A3. Wayne, PA: Clinical and Laboratory Standards Institute, 2014
20. Evaluation of Detection Capability for Clinical Laboratory Measurement Procedures; Approved Guideline—Second Edition. CLSI document EP17−A2. Wayne, PA: Clinical and Laboratory Standards Institute, 2012
21. Evaluation of the Linearity of Quantitative Measurement Procedures: A Statistical Approach: Approved Guideline. CLSI document EP6-A. Wayne, PA: Clinical and Laboratory Standards Institute, 2003
22. Interference Testing in Clinical Chemistry; Approved Guideline−Second Edition. CLSI document EP7-A2. Wayne, PA: Clinical and Laboratory Standards Institute, 2005
23. User Protocol for Evaluation of Qualitative Test Performance−Second Edition. CLSI document EP12-A2. Wayne, PA: Clinical and Laboratory Standards Institute, 2008
24. Estimation of Total Analytical Error for Clinical Laboratory Methods: Approved Guideline. CLSI document EP21- A. Wayne, PA: Clinical and Laboratory Standards Institute, 2003
25. Assessment of the Diagnostic Accuracy of Laboratory Tests Using Receiver Operating Characteristic Curves; Approved Guideline−Second Edition. CLSI document EP24−A2. Wayne, PA: Clinical and Laboratory Standards Institute, 2011
26. Defining, Establishing, Verifying Reference Intervals in the Clinical Laboratory; Approved Guideline−Third Edition. CLSI document C28−A3. Wayne, PA: Clinical and Laboratory Standards Institute, 2008
27. Agresti A: Categorical Data Analysis. New York, NY: Wiley, 1990. pp. 422−425
28. Altman DG: Categorizing continuous variables. Br J Cancer 1991, 64:975

29. Zweig MH, Campbell G: Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clin Chem 1993, 39:561−577
30. Zhou XH, Obuchowski NA, McClish DK: Statistical Methods in Diagnostic Medicine−Second Edition. New York, NY: Wiley, 2011
31. Pepe MS: The Statistical Evaluation of Medical Tests for Classification and Prediction. Oxford: Oxford University Press, 2003
32. Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JHP, Bossuyt PMM: Empirical evidence of design-related bias in studies of diagnostic tests. JAMA 1999, 282:1061−1066
33. Begg CB: Biases in assessment of diagnostic tests. Stat Med 1987, 6: 411−423
34. Ransohoff DF, Feinstein AR: Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. N Engl J Med 1987, 299: 926−930
35. Begg CB, Greenes RA: Assessment of diagnostic tests when disease verification is subject to selection bias. Biometrics 1983, 39:207−215
36. Begg CB, Greentest RA, Iglewicz B: The influence of uninterpretability on the assessment of diagnostic test. J Chronic Dis 1986, 39: 575−584
37. Hadgu A: The discrepancy in discrepant analysis. Lancet 1996, 348: 592−593
38. Hadgu A: Bias in the evaluation of DNA-amplification tests for detecting chlamydia trachomatis. Stat Med 1997, 16:1391−1399
39. Campbell G, Pennello G, Yue L: Missing data in the regulation of medical devices. J Biopharm Stat 2011, 21:180−195
40. National Academy of Sciences: The Prevention and Treatment of Missing Data in Clinical Trials: Panel on Handling Missing Data in Clinical Trials, National Research Council. Washington, DC: National Academies Press, 2010
41. Little RJ, D'Agostino R, Cohen ML, Dickersin K, Emerson SS, Farrar JT, Frangakis C, Hogan JW, Molenberghs G, Murphy SA, Neaton JD, Rotnitzky A, Scharfstein D, Shih WJ, Siegel JP, Stern H: The prevention and treatment of missing data in clinical trials. N Engl J Med 2012, 367:1355−1360
42. Biswas B: Assessing Agreement for Diagnostic Devices. Washington, DC: FDA-Industry Statistics Workshop, 2006
43. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, Lijmer JG, Moher D, Rennie D, de Vet HC: Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Clin Chem 2003, 49:1−6
44. Pepe MS, Feng Z, Janes H, Bossuytt PM, Potter JD: Pivotal evaluation of the accuracy of a biomarker used for classification or prediction: standards for study design. J Natl Cancer Inst 2008, 100: 1432−1438
45. Pepe MS, Etzioni R, Feng Z, Potter JD, Thompson ML, Thomquist M, Winget M, Yasui Y: Phases of biomarker development for early detection of cancer. J Natl Cancer Inst 2001, 93: 1054−1061
46. US Food and Drug Administration. Guidance for industry, clinical investigators, institutional review boards and food and drug administration staff: design considerations for pivotal clinical investigations for medical devices. Silver Spring, MD, US FDA, 2013. Available at http://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/UCM373766.pdf (accessed November 3, 2014)
47. US Food and Drug Administration. Guidance for industry and FDA staff: statistical guidance on reporting results from studies evaluating diagnostic tests. Silver Spring, MD, US FDA, 2007. Available at http://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm071287.pdf (accessed January 8, 2016)
48. US Food and Drug Administration. Guidance for industry and Food and Drug Administration staff: establishing the performance characteristics of in vitro diagnostic devices for the detection or detection and differentiation of human papillomaviruses. Silver Spring, MD, US FDA, 2011. Available at http://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/UCM181511.pdf (accessed May 14, 2015).
49. Simon RM, Paik S, Hayes DF: Use of archived specimens in evaluation of prognostic and predictive biomarkers. J Natl Cancer Inst 2009, 101:1446−1452
50. Walter SD, Macaskill P, Lord SJ, Irwig L: Effect of dependent errors in the assessment of diagnostic or screening test accuracy when the reference standard is imperfect. Stat Med 2012, 31:1129−1138
51. Walter SD: Estimation of test sensitivity and specificity when disease confirmation is limited to positive results. Epidemiology 1999, 10: 67−72
52. Biggerstaff BJ: Comparing diagnostic tests: a simple graphic using likelihood ratio. Stat Med 2000, 19:649−663
53. Schatzkin A, Connor RJ, Taylor PR, Bunnag B: Comparing new and old screening tests when a reference procedure cannot be performed on all screeners. Am J Epidemiol 1987, 125:672−678
54. Feinstein AR, Cicchetti DV: High agreement but low kappa, I: the problems of two paradoxes. J Clin Epidemiol 1990, 43:543−549
55. Feinstein AR, Cicchetti DV: High agreement but low kappa, II: resolving the paradoxes. J Clin Epidemiol 1990, 43:551−558
56. Fleiss JL: Statistical Methods for Rates and Proportions. ed 2. New York, NY, Wiley, 1981
57. Leisenring W, Alonzo TA, Pepe MS: Comparison of predictive values of binary medical diagnostic tests for paired designs. Biometrics 2000, 56:345−351
58. Alonzo TA, Pepe MS, Moskowitz CS: Sample size calculations for comparative studies of medical tests for detecting presence of disease. Stat Med 2002, 21:835−852
59. Connor RJ: Sample size for testing differences in proportions for the paired-sample. Biometrics 1987, 43:207−211
60. Nam J: Power and sample size requirements for non-inferiority in studies comparing two matched proportions where the events are correlated. Comput Stat Data Anal 2011, 55:2880−2887
61. Newcombe RG: Two-sided confidence intervals for the single proportion: comparison of seven methods. Stat Med 1998, 17:857−858
62. Agresti A, Coull BA: Approximate is better than "exact" for interval estimation of binomial proportions. Am Stat 1998, 52:119−126
63. Krõlov K, Frolova J, Tudoran O, Suhorutsenko J, Lehto T, Sibul H, Mäger I, Laanpere M, Tulp I, Langel Ü: Sensitive and rapid detection of chlamydia trachomatis by recombinase polymerase amplification directly from urine samples. J Mol Diagn 2014, 16:127−135
64. Mercaldo ND, Lau KF, Zhou XH: Confidence intervals for predictive values with an emphasis to case−control studies. Stat Med 2007, 26: 2170−2183