Evaluation of diagnostic tests for infectious diseases: general principles

The TDR Diagnostics Evaluation Expert Panel

I. INTRODUCTION

A diagnostic test for an infectious agent can be used to demonstrate the presence or absence of infection, or to detect evidence of a previous infection (for example, the presence of antibodies). Demonstrating the presence of the infecting organism, or a surrogate marker of infection, is often crucial for effective clinical management and for selecting other appropriate disease control activities such as contact tracing. To be useful, diagnostic methods must be accurate, simple and affordable for the population for which they are intended. They must also provide a result in time to institute effective control measures, particularly treatment. For some infections, early diagnosis and treatment can have an important role in preventing the development of long-term complications or in interrupting transmission of the infectious agent. In a broader context, diagnostic tests can have multiple uses, including: patient management, especially when clinical symptoms are not specific for a particular infection (as is often the case); screening for asymptomatic infections; surveillance; epidemiological studies (for example, rapid assessments of disease burden or outbreak investigations); evaluating the effectiveness of interventions, including verification of elimination; and detecting infections with markers of drug resistance.

Recent technological developments have led to the proliferation of new, rapid diagnostic tests that hold promise for the improved management and control of infectious diseases. Whether these tests are useful in a given setting and, if so, which test is most appropriate are questions that can be answered only through evaluations in the appropriate laboratory, clinical or field settings.

Many variables can influence the performance of tests in different settings. These include differences in the characteristics of the population or the infectious agent, including the infection prevalence and genetic variation of the pathogen or host, as well as the test methodology — for example, the use of recombinant or native antigen or antibody, whether the test is manual or automatic, the physical format of the test and local diagnostic practice and skills. Therefore, wherever possible, test evaluations should be performed under the range of conditions in which they are likely to be used in practice. In some situations, such evaluations can be facilitated through multi-centre trials.

Lack of resources and expertise limit the ability of many developing countries to perform adequate evaluations of diagnostic tests, and many new tests are marketed directly to end-users who lack the ability to assess their performance. The onus is therefore on those who perform the evaluations to ensure that the quality of the methods and the documentation used is such that the findings add usefully to the pool of knowledge on which others can draw. The Standards for Reporting of Diagnostic Accuracy (STARD) initiative has developed a sequenced checklist to help to ensure that all relevant information is included when the results of studies on diagnostic accuracy are reported[1–4] (APPENDIX 1).

Evaluations of diagnostic tests must be planned with respect to their use for a clearly defined purpose, carefully and systematically executed, and must be reported in a way that allows the reader to understand the study methods and the limitations involved and to interpret the results correctly. This will help to avoid the financial and human costs associated with incorrect diagnoses, which can include poor patient care, unnecessary complications, suffering and, in some circumstances, even death.

This document is concerned with general principles in the design and conduct of trials to evaluate diagnostic tests. It is not a detailed operational manual and should be used alongside detailed descriptions of statistical methods, clinical trial guides and other reference materials given in the reference list.

The goals of this document are to facilitate the setting of appropriate standards for test evaluation; to provide best-practice guidelines for assessing the performance and operational characteristics of diagnostic tests for infectious diseases in populations in which the tests are intended to be used; to help those designing evaluations at all levels, from test manufacturers to end-users; and to facilitate critical review of published and unpublished evaluations, with a view to selecting or approving tests that have been appropriately evaluated and shown to meet defined performance targets. The target audience for this document includes institutions and research groups that are planning trials of diagnostic tests; organizations that fund or conduct trials of diagnostic tests; agencies responsible for the procurement of diagnostic tests; diagnostic test manufacturers; and regulatory authorities.

II. CHARACTERISTICS ASSESSED IN EVALUATIONS OF DIAGNOSTIC ACCURACY

1. Performance characteristics

The basic performance characteristics of a test designed to distinguish infected from uninfected individuals are sensitivity, that is, the probability that a truly infected individual will test positive, and specificity, that

is, the probability that a truly uninfected individual will test negative. These measures are usually expressed as a percentage.

Sensitivity and specificity are usually determined against a reference standard test, sometimes referred to as a 'gold standard' test, that is used to identify which subjects are truly infected and which are uninfected. Errors in measuring the sensitivity and specificity of a test will arise if the 'gold standard' test itself does not have 100% sensitivity and 100% specificity, which is not infrequently the case. Evaluating a diagnostic test is particularly challenging when there is no recognized reference standard test.

Two other important measures of test performance are positive predictive value (PPV), the probability that those testing positive by the test are truly infected, and negative predictive value (NPV), the probability that those testing negative by the test are truly uninfected. Both of these measures are often expressed as percentages. PPV and NPV depend not only on the sensitivity and specificity of the test, but also on the prevalence of infection in the population studied (BOX 1). The reproducibility of a test is an assessment of the extent to which the same tester achieves the same results on repeated testing of the same samples, or the extent to which different testers achieve the same results on the same samples, and is measured by the percentage of times the same results are obtained when the test is used repeatedly on the same specimens. Reproducibility can therefore be measured between operators or with the same operator, or using different lots of the same test reagent. The accuracy of a test is sometimes used as an overall measure of its performance and is defined as the percentage of individuals for whom both the test and the reference standard give the same result (that is, the percentage of individuals whom both tests classify as infected or uninfected). Note that the use of this measure of diagnostic accuracy is of limited value and is often difficult to interpret, as it depends on sensitivity, specificity and the prevalence of infection.

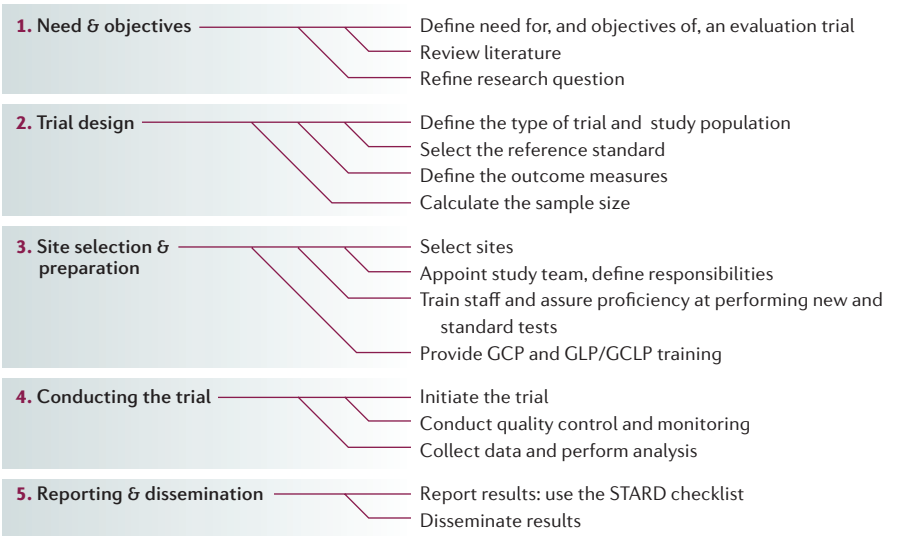


Figure 1 | **Essential elements in designing diagnostic test evaluations.** GCP, good clinical practice; GCLP, good clinical laboratory practice; GLP, good laboratory practice; STARD, standards for reporting of diagnostic accuracy. See Section III, 2.13.

2. Operational characteristics

Operational characteristics include the time taken to perform the test, its technical simplicity or ease of use, user acceptability and the stability of the test under user conditions. The ease of use will depend on the ease of acquiring and maintaining the equipment required to perform the test, how difficult it is to train staff to use the test and to interpret the results of the test correctly, and the stability of the test under the expected conditions of use. All of these characteristics are important for determining the settings in which a diagnostic test can be used and the level of staff training required. Information on test stability — how tests can be stored in peripheral healthcare settings and for how long — is crucial for decisions on procurement.

III. ESSENTIAL ELEMENTS IN THE DESIGN OF DIAGNOSTIC TEST EVALUATIONS

The design of a study is likely to be greatly improved if this process is approached systematically along the lines outlined in FIGURE 1.

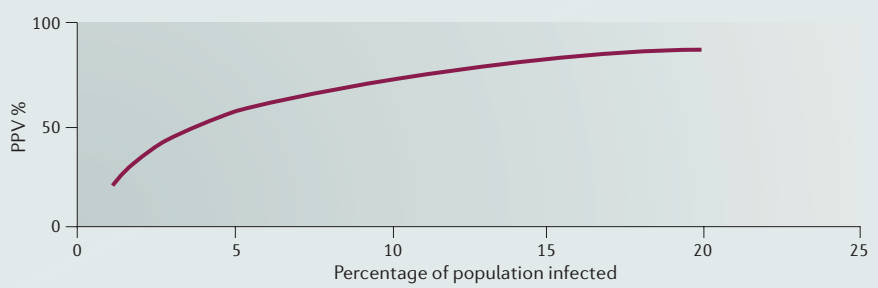
1. Defining the need for a trial and the trial objectives

Before conducting a trial, the specific need for the trial should be assessed. The purpose of the study and the degree to which the outcome is likely to contribute to improved health outcomes and/or further knowledge about the performance of the test should be specified.

First, the problem must be identified. Examples include the need for a screening test because an infection is asymptomatic but undetected infection can cause serious complications, as is the case for screening pregnant women to prevent congenital syphilis; the need for a rapid point-of-care test because of low patient return for results of laboratory-based tests that require return visits; and the need for a test that can be performed on specimens obtained through less-invasive procedures, such as a blood test instead of a lumbar puncture for determining the stage of disease in African trypanosomiasis.

Box 1 | **Dependence of PPV on prevalence**

The positive predictive value (PPV) of a test will depend not only on the sensitivity of the test but also on the prevalence of the condition within the population being tested. The figure below shows how the positive predictive value for a test with 96% sensitivity varies according to the prevalence of infection in the population.



The purpose for which the new test is designed, for example, whether the test is to be used for case management, to screen asymptomatic infections, for surveillance or to verify elimination, must also be defined. This should include defining the objectives of the evaluation — both the overall objective, for example, improving the quality of diagnosis for patient management or surveillance, and the specific objectives, for example, assessing test performance, acceptability, the impact on patient care or the prevention of complications.

Finally, the relevant literature must be reviewed and the research question refined. Planning a study should include a comprehensive review of the relevant literature for the diagnostic test under evaluation and other relevant tests for the infection under study, including an assessment of the strengths and limitations of previous studies. Where possible, manufacturers' clinical and analytical data for the new test should be assessed. The outcome of a review of previous work should inform assessment of the need for another trial and the specific areas in which further information is needed. As a consequence of the review, the research question might be refined.

2. General considerations in the design of evaluation trials

The design of the evaluation trial will depend on its purpose and the population for which the test is intended. See REF. 5 for a general discussion of the factors to take into account when planning field trials. In general, a diagnostic test should be evaluated using methods and equipment that are appropriate for that purpose. The staff performing the evaluation should be appropriately trained so that they are proficient in performing the test being evaluated and the comparator tests.

2.1. Defining the study population. There should be clear specification of the eventual target population in which the diagnostic test will be used. Defining the target population must take into account the probable purpose of the test. For example, will it replace an existing test, will it be used as a triage instrument to identify those in need of further investigation, or will it be used as an additional test in a diagnostic strategy for case finding or for screening asymptomatic individuals? The actions that are to be guided by the use of the test, such as starting or withholding treatment, must also be considered. It is of little value to evaluate tests that are unlikely to be affordable or accessible to the target population, or which yield results that are unlikely to influence patient care or public health practice.

2.2. Subjects to be included in the study. Two common circumstances in which diagnostic tests are deployed are:

a Screening people presenting to a clinic who have symptoms that might be caused by the infection to identify those who are truly infected (for example, persons presenting with a fever that might be caused by malaria).

b Distinguishing infected people from non-infected people in a population, irrespective of whether or not they have any symptoms that might be characteristic of the infection.

Generally, in situation (a), tests with high sensitivity will be required so that a high proportion of all truly infected patients are identified for treatment. In situation (b), if the infection is rare, high specificity will be required or else a high proportion of those who test positive could be false positives (that is, the test will have a poor PPV). In either circumstance it is necessary to identify a group of truly infected and truly uninfected individuals to assess sensitivity and specificity, respectively.

For situation (a), a common design for an evaluation study is to enroll consecutive subjects who are clinically suspected of having the target condition. The suspicion of infection can be based on presenting symptoms or on a referral by another healthcare professional. These participants then undergo the test under evaluation as well as the reference standard test. In studies in which only a small proportion of those tested are likely to be infected, all subjects can be subjected to the reference standard test first. All positives and only a random sample of test negatives can then be subjected to the test under evaluation. This can lead to more efficient use of resources if the target condition is rare.

Tests can sometimes be evaluated using stored specimens collected from those with known infection status. Such studies are rapid and can be of great value but there is a risk that they can lead to inflated estimates of diagnostic accuracy (when the stored samples have been collected from the 'sickest of the sick' and the 'healthiest of the well'). The estimate of specificity can also be biased if, for example, the negative samples relate only to a group with one alternative condition, rather than a group including the full range of conditions that can present with symptoms that are similar to the infection under study.

Box 2 | **Multi-centre studies**

Advantages
- Larger sample size
- Representative of more than one population so findings are more generally applicable

Disadvantages
- Greater expense
- Quality control and coordination more difficult
- Site conditions might not be comparable

2.3. The study setting. The setting where patients or specimens will be recruited and where the evaluation will be conducted should be defined. This might be in a clinic or laboratory, a remote health post or a hospital. Tests will probably perform differently in a primary care setting compared with a secondary or tertiary care setting. The spectrum of endemic infections and the range of other conditions observed can vary from setting to setting, depending on the referral mechanism. Other factors that can affect test performance and differ between sites include climate, host genetics and the local strains of pathogens. Because the test characteristics can vary in different settings, it is often valuable to consider conducting multi-centre studies. Some of the advantages and disadvantages of multi-centre studies are shown in BOX 2.

2.4. Retrospective and prospective evaluations. Diagnostic evaluations can be performed both retrospectively, using well-characterized archived specimens, and prospectively, using fresh specimens. The choice depends on the availability of appropriate specimens and whether the research question can be answered wholly or in part using archived specimens. Some advantages and disadvantages of using archived specimens are shown in BOX 3.

2.5. Eligibility criteria. The eligibility criteria are generally defined by the choice of the target population, but additional exclusion criteria can be used for reasons of safety or feasibility. The researcher must consider, for example, whether or not patients with co-morbidity or other conditions likely to influence the study results will be excluded. For infectious diseases, additional exclusion

Box 3 | **Using archived specimens**

Advantages
- Convenience
- Speed
- Economy

Disadvantages
- Specimen quality can be affected by storage
- Patient information (e.g. age, sex and severity of symptoms) might be limited or not available
- Specific informed consent for such testing might not have been given at the time of specimen collection, so informed consent might need to be obtained or, if this is not possible, personal identifiers and patient information should be removed from specimens for testing

criteria might include recent use of antibiotics or other treatments. Such exclusions can make results easier to interpret but might also limit their ability to be applied generally to populations in which the test might be used in practice.

2.6. Sampling. The study group can consist of all subjects who satisfy the criteria for inclusion and are not disqualified by one or more of the exclusion criteria. In this case, a consecutive series of subjects is often included. Alternatively, the study group can be a sub-selection, for example, only those who test negative by the reference test. However, this can lead to biased estimates if the sample is not truly random.

2.7. Selecting the reference standard test. Where possible, all tests under evaluation should be compared with a reference (gold) standard. The choice of an appropriate reference standard is crucial for the legitimacy of the comparison. For example, a serological assay should not usually be compared with an assay that detects a microorganism directly, and clinically defined reference standards are not usually appropriate when clinical presentation is not sensitive or specific. Non-commercial or 'in-house' reference standards are legitimate only if they have been adequately validated. Sometimes, composite reference standards might have to be used in the absence of a single suitable reference standard. Results from two or more assays can be combined to produce a composite reference standard[6]. For example, if there are two possible 'gold standard' tests, both of which have high specificity but poorer sensitivity, then positives can be defined as samples that test positive by either test. In other circumstances, positives can be defined as those that test positive by both tests, negatives as those that test negative by both tests, and others omitted from the evaluation as indeterminate.

New tests under evaluation that are more sensitive than the existing reference standard usually require a composite reference standard. If a reference standard is not available and a composite standard cannot be constructed, an appropriate approach might be to report the levels of agreement between different tests, that is, positive by both or negative by both.

2.8. Evaluating more than one test. If more than one test is being evaluated, the evaluations can be sequential or simultaneous. The advantages and disadvantages of conducting simultaneous comparisons of several tests are listed in BOX 4.

2.9. Defining the outcome measures. The outcomes of the evaluation, such as performance characteristics, should be clearly defined. Evaluations should always include 95% confidence intervals for sensitivity and specificity (TABLE 1; Section 2.10).

In the absence of a reference standard, the performance of the test under evaluation can be compared to an existing test using a 2×2 table, which shows how the samples were classified by each test. The values for percentage agreement positive, percentage agreement negative, percentage negative by test 1 and positive by test 2, and percentage positive by test 1 and negative by test 2 can be derived from such a table. In addition, for prospective evaluations PPV and NPV can be used. These values will depend on the prevalence of the infection in the studied population.

In cases where the interpretation of test results is subjective, such as visual reading of a dipstick test result, an important outcome measure is assessment of the agreement between two or more independent readers.

2.10. Calculating the sample size. The key question to be addressed before embarking on a study, and the question that is often hardest to answer, is what level of performance is required of the test. The levels that might be acceptable in one setting might be inappropriate in another. The indications for performing the test can vary. The level and availability of healthcare resources and disease prevalence all have a bearing on setting the acceptable performance of a test.

Increasing the sample size reduces the uncertainty regarding the estimates of sensitivity and specificity (the extent of this uncertainty is summarized by the confidence interval). The narrower the confidence interval, the greater the precision of the estimate. A 95% confidence interval is often used — that is, we can be 95% certain that the interval contains the true values of sensitivity (or specificity). The formula for calculating the 95% confidence interval is given by equation 1

$$p \pm 1.96 \times \sqrt{\frac{p(1-p)}{n}} \qquad (1)$$

where p = sensitivity (or specificity) measured as a proportion (not a percentage) and n = number of samples from infected people (or, for specificity, from uninfected people).

Box 4 | **Advantages and disadvantages of simultaneous comparisons**

Advantages
- Provide 'head to head' comparisons for two or more tests using a single reference standard and the same patient population
- Speed: results are available sooner than if conducting sequential trials
- More cost-effective

Disadvantages
- Can be difficult to interpret results for several tests independently, as blinding is not usually possible (that is, the results of other tests on the same samples or individuals might be known to the testers)
- Complicates the design of the evaluation, for example, randomization of the order in which specimens for the different tests are collected and assessed
- Specimen quality can be compromised with increasing numbers of specimens collected, especially with specimens collected from limited anatomical sites, such as urethral swabs or finger-prick blood specimens
- The collection of multiple specimens might not be acceptable to patients

Table 1 | **A 2 × 2 table to evaluate test performance**

Test under evaluation	Reference standard test		Total
	Positive	Negative	
Positive	a	b	a + b
Negative	c	d	c + d
Total	a + c	b + d	

Test sensitivity = a/(a + c); test specificity = d/(b + d); PPV = a/(a + b); NPV = d/(c + d).
a = true positive, b = false positive; c = false negative; d = true negative

As an example of how confidence intervals are calculated, suppose 97 samples are positive by the 'gold standard' test and 90 of these are positive by the test under evaluation, then the sensitivity of the test is estimated by $p = 90/97 = 0.928$ and the confidence interval, using the formula above, is given in equation 2.

$$0.928 \pm 1.96 \times \sqrt{\frac{0.928(1 - 0.928)}{97}} = 0.928 \pm 0.051 = 0.877–0.979 \quad (2)$$

That is, we are 95% sure that the interval 87.7% to 97.9% contains the true sensitivity of the test under evaluation.

In considering sample size, it is important to consider the desired precision with which the sensitivity (or specificity) of the test is to be measured. To do this, we must first make a rough estimate of what we expect the sensitivity (or specificity) to be. So, if we suspect the sensitivity (or specificity) of the test under evaluation is approximately p (for example, 0.8 (80%)) and we wish to measure the sensitivity (or specificity) to within ± x (where x is expressed as a proportion rather than a percentage; for example, 0.10 rather than 10%) then we would choose n so that the confidence interval is ± x (for example ± 10%). This is shown in equations 3–5.

$$\sqrt{\frac{p(1 - p)}{n}} \leq x \quad (3)$$

which translates to:

$$n \geq \frac{(1.96)^2\, p(1 - p)}{x^2} \quad (4)$$

Thus, if $p = 0.80$ and $x = 0.10$, then

$$n \geq \frac{(1.96)^2\, 0.8(1 - 0.8)}{(0.1)^2} = 61.5 \quad (5)$$

Therefore, to measure the sensitivity to within ± 10% we require at least 62 samples that are positive by the 'gold standard' test.

TABLE 2 shows the relationship between sample size and 95% confidence interval for various estimated sensitivities and specificities. For example, if we estimate that the sensitivity of a new test is 80% and we want the confidence interval to be ± 6%, we will need to recruit, or have archived specimens from, 170 infected study subjects by the reference standard test. If the prevalence of infection in the study population is 10%, then there will be 10 infected subjects per 100 patients seen at the clinic. So, to have 170 infected subjects, we will need to recruit 1,700 patients (100/10 × 170).

In determining the sample size, allowance must also be made for patients who do not meet the inclusion criteria and the percentage who are likely to refuse to participate in the study.

If, when the study begins, it is not possible to estimate in advance what the sensitivity or specificity will be, then the safest option for the calculation of sample size is to assume these will be 50% (as this results in the largest sample size). Alternatively, sometimes it will be useful to conduct a pilot survey to estimate the prevalence of infection and to obtain a preliminary estimate of sensitivity and specificity. In such a study, the feasibility of the proposed study procedures can also be evaluated.

In some circumstances it might be possible to state the minimal acceptable sensitivity (or specificity) for the intended application of the test. So, if it is suspected that the sensitivity (or specificity) of the test under evaluation is p (for example, 80%) but it is considered that $p_0 = 70\%$ is the minimum acceptable sensitivity (or specificity), then n might be chosen so that the lower limit of the confidence interval is likely to exceed p_0. With the test requirement formulated in this way the sample size formula is given by equation 6:

$$n = (1.96 + 1.28)^2 \frac{p(1 - p)}{(p - p_0)^2} \quad (6)$$

For example, if it is anticipated that the sensitivity of a new test is 80% and to be acceptable for use in a given setting it must be at least 70%, then it will be necessary to recruit, or have archived specimens from, 168 infected study subjects. If the prevalence of infection in the study population is 10%, then it will be necessary to recruit a total sample of 1,680 (168/0.10), to ensure 168 infected individuals.

Details of methods for calculating sample size for diagnostic trials are available in REFS 5,7,8.

2.11. Evaluating reproducibility. The reproducibility of a test is a measure of the closeness of agreement between test results when the conditions for testing or measurement change. For example, reproducibility can be measured between operators (inter- and

Table 2 | **Relationship between sample size and 95% confidence interval**

Number of infected (non-infected) subjects required*	Estimated test sensitivity (or specificity)‡					
	50%	60%	70%	80%	90%	95%
50	13.9%	13.6%	12.7%	11.1%	8.3%	–
100	9.8%	9.6%	9.0%	7.8%	5.9%	4.3%
150	8.0%	7.8%	7.3%	6.4%	4.8%	3.5%
200	6.9%	6.8%	6.4%	5.5%	4.2%	3.0%
500	4.4%	4.3%	4.0%	3.5%	2.6%	1.9%
1,000	3.1%	3.0%	2.8%	2.5%	1.9%	1.4%

*As defined by the reference standard test. ‡95% confidence interval around the estimated sensitivity (+/– value in table).

Glossary

Accuracy
The percentage of correct results obtained by the test under evaluation compared with the results of a reference or 'gold standard' test. Usually expressed as the number of correct results divided by the total number of results, multiplied by 100.

Blinding
Interpreting a test result without knowledge of a patient's condition or previous test results.

Confidence interval
The confidence interval quantifies the uncertainty in measurement; usually reported as the 95% confidence interval, the range that we can be 95% certain covers the true value.

Negative predictive value (NPV)
The probability that a negative result accurately indicates the absence of infection.

Positive predictive value (PPV)
The probability that a positive result accurately indicates the presence of infection.

Prevalence
The proportion of a given population with an infection at a given time.

Proficiency panel
A collection of six or more mock or true specimens with positive and negative results for a particular test, used to ascertain the proficiency of the technologist in performing the test.

Quality assurance (QA)
An ongoing process of monitoring a system for reproducibility or reliability of results, with which corrective action can be instituted if standards are not met.

Reference standard
The best available approximation of a true result, generally indicating a test method that is currently accepted as reasonably, but not necessarily, 100% accurate. It is used as the reference method for assessing the performance characteristics of another test method.

Reproducibility
A measure of the extent to which replicate analyses using identical procedures agree with each other.

Sensitivity
The probability (percentage) that patients with the infection (determined by the result of the reference or 'gold standard' test) will have a positive result using the test under evaluation.

Specificity
The probability (percentage) that patients without the infection (determined by the result of the reference or 'gold standard' test) will have a negative result using the test under evaluation.

Tests
Any method for obtaining additional information regarding a patient's health status.

intra-observer reproducibility), between different test sites, using different instruments, between different kit lots (lot-to-lot reproducibility) or on different days (run-to-run and within-run reproducibility). The Kappa statistic[9] provides a useful measure of agreement between test types or lots, and between users. This statistic allows the measurement of agreement between sets of observations or test results above that expected by chance alone.

When test results are dichotomous (that is, either positive or negative), these characteristics are usually assessed in the following ways — operator-dependent reproducibility (especially important for tests for which the interpretation of results is subjective), in which the same lot of tests is assessed by two operators using the same evaluation panel but blinded to each other's results, and test-dependent reproducibility, which includes lot-to-lot variability, that is, the same operator evaluates different lots of diagnostic product using the same evaluation panel, and run-to-run variability, that is, the same operator evaluates the test several times using the same evaluation panel.

The repeatability of the test results refers to the closeness of the test results when no conditions of measurement change. The extent to which a test will produce the same result when used on the same specimen in identical circumstances (repeatability) should be distinguished from operator-related issues affecting reproducibility, which might be improved by further training.

The study protocol should describe how the reproducibility of the test will be measured and what aspect of reproducibility is being evaluated. This should include a description of the factors that are held constant, for example, reagent lots, instruments, calibration and/or quality-control methods. Reproducibility testing should be conducted in a blinded fashion, that is testers should not know the results obtained previously.

The size of the evaluation panel for reproducibility studies should be dictated by the degree of precision needed for the relevant clinical indication. The panel should include at least one positive and one negative control, and, if appropriate, two or three different operators, with the samples evaluated on three different days. In multi-centre studies, reproducibility should be assessed at each centre and between centres.

As well as measuring the extent to which there is reproducibility in the assessment of strong positive results, it is important to include assessment of reproducibility using weak positive and borderline negative samples if these might be important for clinical decision making.

2.12. Evaluating operational characteristics. An evaluation of a test can also include an assessment of its operational characteristics and cost-effectiveness. The latter is not considered in this document. Some operational characteristics, such as simplicity, acceptability of the test to users, the robustness of the test under different storage conditions and the clarity of instructions, are qualitative and subjective, but assessment of these can be crucial for decisions regarding the suitability of a test for a specific setting. In particular, the robustness of the test under different storage conditions is an area of concern for tests that will be used in remote settings.

Diagnostic tests can contain biological or chemical reagents that are heat-labile and might be affected by moisture, making the shelf-life of the test dependent on the temperature and humidity at which it is stored. Many commercially available *in vitro* diagnostic tests are recommended to be stored between 4°C and 30°C and are sealed in moisture-proof packaging. The specified shelf-life is based on the assumption that these conditions are maintained. Transport and operational conditions in the tropics commonly exceed 30°C, especially for point-of-care tests used in remote areas. Exposure to humidity can occur during delays between opening of the moisture-proof packaging and performance of the test procedure.

During evaluation of diagnostic tests, it is essential to inspect test kits for signs of damage caused by heat or humidity, and to record the conditions under which the tests have been stored and transported. These conditions should be taken into account when interpreting the results. A product dossier of test characteristics, including heat-stability data, should be available from the manufacturer of the diagnostic test. This will assist in extrapolating the results obtained under trial conditions to the results expected if the test kits had been stored under the anticipated operational conditions.

If there is uncertainty about the test stability, storage outside the manufacturer's recommendations is expected during operational use or there are insufficient data on temperature stability, the addition of thermal-stability testing to the trial protocol should be considered. Tests can be stored in an incubator at temperatures near the likely maximum in the field (for example, 40°C for 2–3 months), then assessed in comparison

Box 5 | **Designing a protocol for an evaluation using archived specimens**

1. Define the target population for the test under evaluation
2. Define the type of specimens that should be included in the evaluation panel
3. Define how appropriate specimens should be selected for the evaluation panel and how the specimens should have been stored
4. Calculate the required sample size
5. Develop a method to remove personal identifiers from the specimens (unless previous consent has been given for this type of work) by assigning a study code to each specimen
6. Define how the specimens will be tested to ensure that the results of the reference standard test will not be known when performing the test under evaluation, and vice versa ('blinding')
7. Define a plan to ensure proficiency in performing the reference standard test
8. Define a plan for quality assurance and external validation of trial results
9. Define where the study protocol needs to be sent for ethics approval (local and other relevant ethics committees)
10. Develop a data-analysis plan, for the calculation of sensitivity, specificity and confidence intervals
11. Define the methods for the dissemination of trial results

with tests stored at the recommended temperature during this period. During field evaluations, periodic comparison of the performance of tests stored at ambient temperature in the field against those stored at recommended temperatures should give an indication of the thermal stability of the test and it might be appropriate to stop the evaluation if the results show substantial deterioration of tests.

2.13. Quality assurance and monitoring. All studies should incorporate quality assurance (QA). Study QA procedures should be established to ensure that the studies are conducted and the data are generated, documented and reported in compliance with good clinical laboratory practice (GCLP). GCLP, rather than good laboratory practice (GLP), is more appropriate for trials that are not being undertaken for registration (see http://www.qualogy.co.uk) or for applicable regulatory requirement purposes. QA should be overseen by an individual who is not a member of the study team.

In the context of an evaluation trial, QA comprises:

- Study quality control (SQC): the crucial element of SQC is the generation of, and adherence to, standard operating procedures (SOPs), which comprise detailed and specific written instructions as to how all aspects of the study are to be conducted[10].
- External quality monitoring (EQM): independent monitoring of quality, which can include site visits conducted by a trained study monitor from outside the study team.
- Study quality improvement (SQI): the process through which deficiencies identified through SQC and EQM are remedied.

QA of laboratory and/or diagnostic testing procedures is also crucial in the day-to-day running of a diagnostic laboratory. Laboratory QA comprises internal quality control (IQC), external quality assessment (EQA) and quality improvement measures. IQC refers to the internal measures taken to ensure that laboratory results are reliable and correct, for example, the existence of SOPs for each test procedure, positive and negative controls for assays, stock management to prevent expired reagents being used, and monitoring of specimen quality. EQA, which is sometimes referred to as proficiency testing, is an external assessment of the laboratory's ability to maintain satisfactory quality, ensured by regular testing of an externally generated panel of specimens. Quality improvement is the process through which deficiencies identified through IQC or EQA are remedied and includes staff-training sessions, recalibration of equipment and assessment of the positive and negative controls used for particular tests.

IV. THE DESIGN OF DIAGNOSTIC EVALUATIONS USING ARCHIVED SPECIMENS

If the test evaluation can be undertaken satisfactorily using archived specimens and a panel of well-characterized specimens is available, a retrospective evaluation can be conducted with both the new test and the reference standard. Although this type of study has the advantages of being rapid and relatively inexpensive compared with a prospective study, it is important to consider several factors that might limit the generalizability of the results, including whether the specimens were collected from a population similar to the population in which the test will be used; what clinical and laboratory results are available to characterize the specimens; whether the specimens have been stored appropriately; and whether there are sufficient numbers of positive and negative specimens to provide an adequate sample size.

The steps involved in designing a protocol for an evaluation using archived specimens are outlined in BOX 5.

Details of this information and the procedures to be followed should be stated in the study protocol. External validation can be performed by sending all positive specimens and a proportion of the negative specimens to another laboratory for testing. Informed consent is usually not required for trials using archived specimens from which personal identifiers have been removed. Some ethics review committees require the investigator to provide information on how the specimens can be

Box 6 | **Designing a protocol for a prospective evaluation**

1. Define the target population for the test under evaluation
2. Develop methods for the recruitment of study participants and informed consent procedures
3. Design study instruments such as data-collection forms and questionnaires
4. Develop plans to pilot study instruments to determine whether they are appropriate
5. Calculate the required sample size
6. Develop a plan for specimen collection, handling, transport and storage
7. Define how the specimens will be tested to ensure blinding of results of the reference standard test from the results of the test under evaluation
8. Define a plan to ensure proficiency in performing the reference standard test
9. Develop a data-collection and data-analysis plan
10. Develop plans to ensure the confidentiality of study data
11. Define a plan for quality assurance and external validation of trial results
12. Define where the study protocol needs to be sent for ethics approval (local and other relevant ethics committees)
13. Define methods for the dissemination of trial results

Box 7 | **Information to be included in the Patient Information and Consent Forms**

- Purpose of the study
- Study procedures and what is required of participants
- Assurance that participation is voluntary
- Statement of the possible discomfort and risks of participation
- Benefits (for example, treatment or care to be offered to those who test positive by the reference standard test)
- Compensation offered for travel and other out-of-pocket expenses
- Safeguards to ensure confidentiality of patient information
- Freedom to refuse to participate and alternatives to participation, and freedom to withdraw from the study at any time without compromise to future care at the facility
- Use of study data and publication of results
- Contact details of a locally accessible person who can answer questions from participants for the duration of the study
- Participant statement to indicate that they understand what was explained to them and they agree to participate by signing the consent form. Illiterate participants can give consent by a thumbprint witnessed by a third party

made anonymous and require assurance that results cannot be traced to individual patients.

V. THE DESIGN OF PROSPECTIVE DIAGNOSTIC EVALUATIONS

The recommended steps in designing a prospective diagnostic evaluation are outlined in BOX 6.

1. Defining the target population for the test under evaluation

The characteristics of the study population should be fully described (see section III, 2.1)

2. Developing methods for the recruitment of study participants and informed consent procedures

Consider the following:

- Who recruits the study subjects? Ideally, this should not be the clinician caring for the participants, as this might influence the participants' decision.
- Who is eligible for enrolment?
- How will informed consent be obtained? (Recruitment of children will require approval from a parent or guardian.)
- Who will answer participants' questions about the study?
- How will confidentiality be assured?

Further information on informed consent can be obtained from REFS 11 & 12.

The Patient Information and Consent Forms should be clear, concise and in a language (read or narrated) that is understandable to the patient. The forms should include the points outlined in BOX 7. An example consent form is shown in APPENDIX 2. Templates are also available from many academic research ethics review committee websites including the WHO Research Ethics Review Committee (http://www.who.int/rpc/research_ethics/en/). If biological specimens are to be stored for future use, this should be specified in a separate section in the consent form and participants should be given the option to refuse to have their specimens stored but still participate in the study.

In general, the only payment to study subjects should be for compensation for transport to the clinic and loss of earnings because of clinic visits related to the study. Payment should never be excessive, such that it might constitute an undue incentive to participate in the study.

Treatment should usually be provided free of charge. Any treatment decisions (if appropriate) should not be based on the results of the test under evaluation but on the reference test. Refusal to participate in the study should not prejudice access to treatment that would normally be accessible.

3. Designing study instruments

Each item on the patient data form should be considered with respect to the stated aims and objectives of the trial. The collection of unnecessary data is a waste of resources and might detract attention from recording the most important data.

When designing data-record forms, it is advisable to review forms from similar trials; allow adequate time to design, translate (and back-translate) and pilot data forms before starting data collection; specify who will complete each form (interviewer or study subject); and specify the QA procedures to ensure data are recorded correctly.

The layout and content of forms should be discussed with the study staff who will be responsible for data management and analysis. The forms should be user-friendly and data should be easily entered into a database. Consider the paper size and colour, the format of records (record books with numbered pages are preferable to loose sheets of paper) and the use of boxes or lines for multiple-choice responses. Questions can allow open or structured responses. Structured responses limit allowable responses to a predefined list, whereas open responses allow freedom to record unanticipated answers, but are more difficult to code and analyse.

It should be ensured that those who will be completing the forms fully understand the forms and know how to complete the forms correctly. Clarity of language is important, particularly when translation might be necessary and so the forms should use simple, uncomplicated language; avoid abbreviations, ambiguous terms and acronyms; avoid unnecessary wording and compound questions; provide definitions; and translate (and back-translate) all of the questions to ensure the correct data items are recorded.

Ensure that a distinction can be made between omitted responses and responses such as 'not applicable'. Where items are to be skipped, the form should contain documentation of the legitimacy of a skipped answer.

4. Develop plans to pilot study instruments

Plans should be developed to determine whether the study instruments, such as questionnaires and data-collection forms, are appropriate. Questions might need to be rephrased to obtain the relevant response. So far as is possible, all aspects of the study should be piloted in a small study so that the methods and procedures can be modified as appropriate. The pilot study also provides the ability to make a preliminary estimate of infection prevalence, which might aid planning the size of the main study.

5. Calculating the required sample size

See Section III, 2.10.

6. Developing a plan for the study logistics

A plan should be developed for safe specimen collection, handling, transport and storage. Consider using pre-printed unique study numbers for forms and specimens (labels should be tested for adherence when samples are frozen, if necessary). Also, develop a flow diagram for specimen handling that can be distributed to laboratory staff.

7. Defining the blinding of results
Specimens will be tested to ensure blinding of results of the reference standard test from the results of the test under evaluation. Most rapid tests require subjective interpretation of the test result. Steps must be taken to ensure that the staff performing the reference test are not aware of the results from the test under evaluation, and vice versa. Also, laboratory staff should not be aware of clinical findings or of the results of other laboratory tests.

It can be difficult to ensure blinding if several tests are being evaluated at the same time. For any repetitively performed procedures, consider randomizing the order in which they are done —for example, if multiple swabs are to be taken, consider applying the tests in random order to different swabs.

8. Defining a QA plan
A QA plan should be developed for quality management of the diagnostic trial. This includes ensuring that the study personnel are proficient in performing both the tests under evaluation and the reference standard test. Before the start of the trial, the laboratory (or whoever is to perform the tests) should be able to demonstrate proficiency in performing the reference standard test(s). The personnel performing the test under evaluation should also demonstrate proficiency at performing and reading this test. The laboratory should subscribe to external proficiency programmes where available. Training records of study personnel should be kept. The QA plan should also include quality management of study data.

9. Developing a plan for data collection and data analysis
Study results entered into workbooks or directly into computer spreadsheets should be checked daily and signed off by the clinic or laboratory supervisor if possible. When entering results into a computer database, consider double data entry to minimize inadvertent errors. All records and study data should be backed up regularly, preferably daily. Review processes for the study database and approval mechanisms for items to be added or deleted should be established. The form of the tables that will be used in the analysis and the statistical methods that will be used in the interpretation of the study results should be drafted before data have been collected to ensure that all the relevant information will be recorded.

10. Developing plans to ensure the confidentiality of study data
All study data should be kept confidential (for example, in a locked cabinet and a password-protected database, with access limited to designated study personnel).

11. Defining a plan for external validation of trial results
See Section III, 2.13.

12. Scientific and ethical review of study protocol
The study protocol should undergo scientific and ethical review by the relevant bodies. Submission documents for ethics approval must follow national or institutional guidelines. As a minimum, the application document for ethics committee approval should contain the information shown in BOX 8. In addition, some ethics committees require protocols to have undergone prior scientific review.

13. Defining methods for the dissemination of trial results
This can involve submitting results for publication to a scientific journal, but most importantly, there should be a plan to inform those responsible for procuring or authorizing tests of the study findings. Appropriate feedback should be given to study participants.

Box 8 | **Information required in the application document for ethics committees**

- Statement of study objectives and rationale
- Description of study methods
- Preliminary evidence of safety and efficacy
- Type/source of patients or samples
- Primary outcome measure
- Follow up of patients
- Sample size plus rationale for proposed size
- Randomization and method of assignment, if applicable
- Risks and benefits for those participating in the study
- Methods to protect patients from harm
- Safeguards for patient privacy and confidentiality
- Benefits expected to be derived from the study
- Alternatives to participation
- Contact details of a locally accessible person who can answer questions from participants for the duration of the study
- Dissemination of study results and any other relevant material

VI. SITE SELECTION AND STUDY PREPARATION

1. Criteria for selection of field sites
The criteria for field-site selection can include:
- Easy access to suitable target populations.
- Adequate prevalence of infection/disease so that sufficient numbers of infected (and uninfected) people can be recruited.
- Availability of suitably trained study personnel (sometimes further training might be required for the purposes of the trial).
- Adequate facilities for conducting the study, for example, space for conducting confidential interviews.
- Good standard of care available for people found to be infected.
- Capacity to store specimens in correct conditions.
- Sufficient data-handling capacity (for example, staff and computers).
- Ability to perform data analysis (on site, if possible).
- Access to good laboratory facilities (if relevant laboratory accreditation schemes exist, and the laboratory is eligible, it should be accredited).
- A mechanism for ethical review and approval of the trial protocol.

2. Site preparation
2.1. Setting up a trial-management system. From the outset of the trial, a quality-management system should be in place. The composition of the trial team should be clearly defined, as should the responsibilities of each team member and trial-management and trial-monitoring procedures.

2.2. Preparing SOPs. SOPs should be prepared for for all clinical and laboratory procedures required in the study protocol (see REF. 10 and BOX 9).

2.3. Training workshops for GCP and GLP/GCLP. Before the trial begins, the study team should be given training on the principles and implementation of GCP and GLP/GCLP[13].

Box 9 | **Elements to be included in SOPs**

- Recruitment of study participants
- Specimen collection, handling, storage and transport
- Preparation of reagents
- How to use test kits and interpret test results, including handling of indeterminate results
- How to perform reference standard tests
- How to monitor and calibrate equipment
- How to identify and correct malfunctions or errors
- Specific instructions on quality assurance procedures
- Record keeping of trial results

2.4. Assurance of proficiency at performing reference standard and tests under evaluation. Before the trial starts, the laboratory should be able to demonstrate proficiency in performing the reference standard tests as well as the tests under evaluation. The laboratory should subscribe to external proficiency programmes. Training records of study personnel should be kept. Training should be provided for performing the test under evaluation using well-characterized positive and negative specimens.

2.5. Piloting and refining study instruments, including the informed consent process. This is essential to ensure the information is understood by study participants and that the questions are appropriate. Translation of the informed consent information sheet into the local language is also essential. Back-translation is desirable to ensure the accuracy of the information provided to the study participants to allow them to make an informed decision whether or not to participate in the study.

VII. CONDUCTING THE EVALUATION

1. General guidelines on the use of test kits

- Note the lot number and expiry date; a kit should not be used beyond the expiry date.
- Ensure correct storage conditions are in place, as stated by the manufacturer. If this is not possible in the field, or cannot be ensured during transport, this should be made clear when the study is reported. If a desiccant is included in the package, the kit should not be used if the desiccant has changed colour.
- Generally, if test kits are stored in a refrigerator, they should be brought to room temperature approximately 30 minutes before use. The use of cold test kits can lead to false-negative results.
- Damaged kits should be discarded.
- Open test kits only when they have reached room temperature, unless otherwise specified.
- Use test kits immediately after opening.
- Reagents from one kit should not be used with those from another kit.
- Tests should be performed exactly as described in the product insert (if available) or any variations must be clearly noted, such as the method of transferring the sample to the kit or the use of venous blood rather than a finger-prick sample.
- It can be useful to evaluate 'off-label' use: this refers to the use of a test for an indication or with a specimen not mentioned in the package insert, for example, self-administered vaginal swabs or pharyngeal swabs. This can be important in defined circumstances, but the fact that it is off-label use must be clearly stated when the results are reported.

2. Biosafety issues

The investigators must comply with national workplace safety guidelines with regard to the safety of clinic and laboratory personnel and the disposal of infectious waste. General guidelines are given in BOX 10.

3. Trial management

3.1. The facility and equipment. Laboratory facilities and equipment should be available and adequately maintained for the work required, for example, suitable work areas, lighting, storage, ventilation and hand-washing facilities should be available. Where field conditions necessitate different standards of operation, these should be clearly stated in the protocol.

3.2. Proficiency of personnel. There are various options for external QA or proficiency programmes for certain infectious diseases such as the College of American Pathologists Inter-laboratory Survey Programs (http://www.cap.org/apps/cap.portal) or the United Kingdom National External Quality Assessment Service (http://www.ukneqas.org.uk). Ongoing records of performance of proficiency panels should be kept to monitor proficiency, especially when there is a change of personnel.

3.3. Changes of study procedures. Any changes to study procedures should be accompanied by changes in the relevant SOPs. Changes to SOPs should be documented, signed off by the responsible supervisor, dated and disseminated to the study team.

4. Quality assurance

There should be arrangements in place (a QA unit or designated person) to ensure that the study is conducted in accordance with the study protocol. A system should be established so that corrective actions suggested to the study team are properly and rapidly implemented.

5. Trial monitoring

There should be regular independent assessment of the laboratory and/or field team performing the evaluations in compliance with the principles of GCP and GLP/GCLP, including both internal and external quality control and QA procedures.

6. Data analysis

The data should be analysed according to the analysis plan after checking, and if necessary correcting, the study data. The sensitivity and specificity of a test can be calculated by comparing the test results to the validated reference test results. They can be displayed in a 2×2 table, as illustrated in TABLE 1. In addition, for prospective trials, the PPV (a/(a + b)) and NPV (d/(c + d)) can be calculated. Inter-observer variability is calculated as the number of tests for which different results are obtained by two independent readers, divided by the number of specimens tested.

Box 10 | **General biosafety guidelines**

- Treat all specimens as potentially infectious
- Wear protective gloves and a laboratory gown while handling specimens
- Do not eat, drink or smoke in the laboratory
- Do not wear open-toe footwear in the laboratory
- Clean up spills with appropriate disinfectants
- Decontaminate all materials with an appropriate disinfectant
- Dispose of all waste, including all clinical material and test kits, using an appropriate method such as placing sharp objects in a biohazard container and disposable materials in sealable waste bags for incineration

VIII. REPORTING AND DISSEMINATING RESULTS

Wherever possible, study participants should be given feedback on study results by, for example, meeting with the study community or having a readily accessible contact person at a clinic to answer specific queries. The results can also be disseminated by publication in peer-reviewed journals or posted on relevant websites. The STARD checklist should be used to guide how a study is reported (APPENDIX 1).

Currently, published studies vary in their attainment of the STARD criteria, often succumbing to common pitfalls[14–16] including inadequate data being used as evidence (including inadequate sample size); bias (for example, by poor selection of study subjects, inappropriate representation of the intended target population, lack of blinding or the use of poor or no reference standards); inadequate description of the characteristics of the study population (for example, parasite density can affect the sensitivity of malaria tests); and evaluations in populations for which the tests are not intended.

IX. CONCLUSIONS

The rapid advances that have been made in molecular biology and molecular methods have led, and continue to lead, to the development of sensitive and specific diagnostic tests, which hold the promise of substantially strengthening our ability to diagnose, treat and control many of the major infectious diseases in developing countries. It is imperative that these new diagnostics are rigorously and properly evaluated in the situations in which they will be deployed in disease control before they are released for general use. A poorly performing diagnostic might not only waste resources but might also impede disease control. The basic procedures described in this article for designing and conducting diagnostic evaluations provide an outline for ensuring the proper evaluation of new diagnostics in laboratory and field trials.

Shabir Banoo is at the Medicines Control Council of South Africa, Pretoria, South Africa.

David Bell is at the Malaria and other Vector-borne and Parasitic Diseases, World Health Organization–Regional Office for the Western Pacific, Manila, Philippines.

Patrick Bossuyt is at the Department of Clinical Epidemiology and Biostatistics, Academic Medical Centre, University of Amsterdam, Amsterdam, The Netherlands.

Alan Herring is at the Veterinary School, University of Bristol, Bristol, UK.

David Mabey is at the Clinical Research Unit, London School of Hygiene and Tropical Medicine, London, UK.

Freddie Poole is at the Division of Microbiology Devices, Center for Biologics Evaluation and Research, Food and Drug Administration, Rockville, USA.

Peter G. Smith is at the Infectious Diseases Epidemiology Unit, London School of Hygiene and Tropical Medicine, London, UK.

N. Sriram is at the Tulip Group of Companies, Goa, India.

Chansuda Wongsrichanalai is at the US Naval Medical Research Unit 2, Jakarta, Indonesia.

Ralf Linke, Rick O'Brien and Mark Perkins are all at the Foundation for Innovative Diagnostics (FIND), Geneva, Switzerland.

Jane Cunningham, Precious Matsoso, Carl Michael Nathanson, Piero Olliaro, Rosanna W. Peeling and Andy Ramsay are all at the UNICEF/UNDP/World Bank/WHO Special Programme for Research & Training in Tropical Diseases (TDR), World Health Organization, Geneva, Switzerland.*

**e-mail: peelingr@who.int*

doi: 10.1038/nrmicro1523

1. Greenhalgh, T. How to read a paper: papers that report diagnostic or screening tests. *Br. Med. J.* **315**, 540–543 (1997).
2. Borriello, S. P. Near-patient microbiological tests. *Br. Med. J.* **319**, 298–301 (1999).
3. Bossuyt, P. M. *et al*. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD Initiative. *Clin. Chem.* **49**, 1–6 (2003).
4. Bossuyt, P. M. *et al*. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Clin. Chem.* **49**, 7–18 (2003).
5. Smith, P. G. & Morrow, R. H., eds *Field Trials of Health Interventions in Developing Countries: A Toolbox*, (Macmillan, London, 1996).
6. Alonzo, T. D & Pepe, M. S. Using a combination of reference tests to assess the accuracy of a new diagnostic test. *Statist. Med.* **18**, 2987–3003 (1999).
7. Pepe, M. S. *Statistical Evaluation of Medical Tests for Classification and Prediction*. (Oxford Univ. Press, 2003).
8. Gardner, M. J. & Altman, D. G. Estimating with confidence. *Br. Med. J.* **296**, 1210–1211 (1988).
9. McGinn, T. *et al.* Tips for learners of evidence-based medicine: 3. Measures of observer variability (kappa statistic). *Can. Med. Assoc. J.* **171**, 1369–1373 (2004).
10. WHO. *Standard Operating Procedures for Clinical Investigators*. UNDP/World Bank/WHO Special Programme for Research and Training in Tropical Diseases (WHO/TDR). TDR/ TDP/SOP99.1 [online] < http://www.who.int/tdr/publications/publications/pdf/sop.pdf > (1999)
11. WHO/Council for International Organizations of Medical Sciences. *International Ethical Guidelines for Biomedical Research Involving Human Subjects* (2002).
12. Nuffield Council on Bioethics. *The Ethics of Research Related to Healthcare in Developing Countries* [online] < http://www.nuffieldbioethics.org/go/ourwork/developingcountries/publication_309.html > (2002)
13. WHO. *Guidelines for Good Laboratory Practice*. UNDP/World Bank/WHO Special Programme for Research and Training in Tropical Diseases TDR/ PRD/GLP/01.2 (WHO/TDR,Geneva, 2001).
14. Delaney, B. C. *et al*. Systematic review of near-patient test evaluations in primary care. *Br. Med. J.* **319**, 824–827 (1999).
15. Reid, M. C., Lachs, M. S. & Feinstein, A. Use of methodological standards in diagnostic test research. Getting better but still not good. *J. Amer. Med. Assoc.* **274**, 645–651 (1995).
16. Small, P. M. & Perkins, M. D. More rigour needed in trials of new diagnostic agents for tuberculosis. *Lancet* **356**, 1048–1049 (2000).

Acknowledgements

We wish to thank Izabela Suder-Dayao for excellent secretarial support, and Robert Ridley and Giorgio Roscigno for support and guidance.

APPENDIX 1 | STANDARDS FOR REPORTING OF DIAGNOSTIC ACCURACY (STARD) CHECKLIST

Section and topic	Item #		On page #
Title/abstract/ keywords	1	Identify the article as a study of diagnostic accuracy (recommended MeSH heading 'sensitivity and specificity').	❑
Introduction	2	State the research questions or study aims, such as estimating the diagnostic accuracy or comparing accuracy between tests or across participant groups.	❑
Methods		Describe:	
Participants	3	The study population: the inclusion and exclusion criteria, the setting and the locations where the data were collected.	❑
	4	Participant recruitment: was the recruitment based on presenting symptoms, results from previous tests, or the fact that the participants had received the index tests or the reference standard?	❑
	5	Participant sampling: was the study population a consecutive series of participants defined by the selection criteria in items 3 and 4? If not, specify how participants were further selected.	❑
	6	Data collection: was data collection planned before (prospective study) or after (retrospective study) the index test and reference standard were performed?	❑
Test methods	7	The reference standard and its rationale.	❑
	8	Technical specifications of the material and methods involved, including how and when the measurements were taken, and/or cite references for the index tests and reference standard.	❑
	9	Definition of, and rationale for, the units, cut offs and/or categories of the results of the index tests and the reference standard.	❑
	10	The number, training and expertise of the persons executing and reading the index tests and the reference standard.	❑
	11	Whether or not the readers of the index tests and reference standard were blind to the results of the other test and describe any other clinical information available to the readers.	❑
Statistical methods	12	Methods for calculating or comparing measures of diagnostic accuracy, and the statistical methods used to quantify uncertainty (e.g. 95% confidence intervals).	❑
	13	Methods for calculating test reproducibility, if done.	❑
Results		Report:	❑
Participants	14	When the study was done, including the start and end dates of recruitment.	❑
	15	The clinical and demographic characteristics of the study population (e.g. age, sex, spectrum of presenting symptoms, co-morbidity, current treatments and recruitment centres).	❑
	16	The number of participants satisfying the criteria for inclusion that did or did not undergo the index tests and/or the reference standard; describe why participants failed to receive either test (a flow diagram is strongly recommended).	❑
Test results	17	Time interval from the index tests to the reference standard, and any treatment administered inbetween.	❑
	18	Distribution of severity of disease (define criteria) in those with the target condition; other diagnoses in participants without the target condition.	❑
	19*	A cross tabulation of the results of the index tests (including indeterminate and missing results) by the results of the reference standard.	❑
	20	Any adverse events from performing the index tests or the reference standard.	❑
Estimates	21	Estimates of diagnostic accuracy and measures of statistical uncertainty (e.g. 95% confidence intervals).	❑
	22	How indeterminate results, missing responses and outliers of the index tests were handled.	❑
	23	Estimates of variability of diagnostic accuracy between subgroups of participants, readers or centres, if done.	❑
	24	Estimates of test reproducibility, if done.	❑
	25	Discuss the clinical applicability of the study findings.	❑

* This entry has been modified from the original.

APPENDIX 2 | SAMPLE INFORMED CONSENT FORM

(A separate patient information sheet containing this information should also be provided)

A | PURPOSE OF THE STUDY

Chlamydial infection is caused by bacteria that are transmitted by sexual intercourse. In women, this infection can cause pelvic pain and, in the long term, increase the risk of infertility. Furthermore, during unprotected sexual intercourse with a man infected with the AIDS virus, a woman infected with chlamydia will have a higher risk of acquiring the AIDS virus than a woman not infected with this bacterium.

To find out whether you have this infection, we need to do some laboratory tests. These tests are expensive and the results are not available the same day. Rapid tests to diagnose chlamydia within 30 minutes are now available but we do not know if they are accurate or reliable. The main purpose of this study is to evaluate a rapid test for the diagnosis of chlamydial infection. We would like to compare the result of these rapid tests with a laboratory-based test to see if they are as accurate as laboratory tests.

B | STUDY PROCEDURES

If you agree to participate in the study, you will be assigned a study number. The doctor or nurse will give you a physical examination and ask you some questions according to standard clinic procedure. He/she will take two samples from your vagina and two samples from your cervix. Your name will not appear on any samples or on the questionnaire. All the samples will be destroyed at the end of the study. If you are diagnosed with chlamydial infection using the standard laboratory tests, you will be treated with antibiotics on your follow-up visit according to normal clinic procedure. You will not be treated according to the rapid test results as we are not yet sure if it is accurate.

C | VOLUNTARY PARTICIPATION

Your decision not to participate in this study will not affect the care you will receive at the clinic in any way. Even if you do agree to become a study participant, you can withdraw from the study at any time (verbally) without affecting the care that you will receive. During the interview, you can choose not to answer any particular question.

D | DISCOMFORT AND RISKS

You may feel a small amount of discomfort or have a small amount of bleeding from the vagina after the pelvic examination and specimen collection.

E | BENEFITS

There will be no immediate benefits from your participation in the study. When the study results are known and if the rapid tests are acceptable in terms of accuracy, everyone who comes to the clinic could benefit from having this test available to diagnose chlamydia and receive the right treatment the same day.

F | COMPENSATION

There will be no monetary compensation for this study, but routine medical consultation and appropriate referral services are available.

G | CONFIDENTIALITY STATEMENT

The records concerning your participation are to be used only for the purpose of this research project. Your name will not be used on any study forms or labels on laboratory specimens or in any report resulting from this study. At the beginning of the study, we will give you a study identification number and this number will be used on the forms and on the laboratory specimens. Any information obtained in connection with this study will be kept strictly confidential. Only members of the study team will have access to information linking your name with your study number.

H | QUESTIONS AND FREEDOM TO WITHDRAW FROM THE STUDY

You can withdraw from the study at any time without affecting your present or future medical care at the clinic. You can contact any of the study personnel if you have questions about the research. (Please give the contact name, address and telephone number of the contact person for each site).

I | RESULTS PUBLICATION

When the researchers have analysed the data, the results and the explanation of its implications will be posted at the clinic for everyone's information.

J | PARTICIPANT STATEMENT

I have been informed verbally and in writing about this study and understand what is involved. I also know whom to contact if I need more information. I understand that confidentiality will be preserved. I understand that I am free to withdraw from the study at any time without affecting the care I normally receive at the clinic. I agree to participate in this study as a volunteer subject and will be given a copy of this information sheet to keep.

_____ _____

Date Name of participant

_____

Signature (or thumb print or cross) of participant

_____ _____

Date Name of witness

_____

Signature of witness

K | INVESTIGATOR'S STATEMENT

I, the undersigned, have defined and explained to the volunteer in a language she understands, the procedures of this study, its aims and the risks and benefits associated with her participation. I have informed the volunteer that confidentiality will be preserved, that she is free to withdraw from the trial at any time without affecting the care she will receive at the clinic. Following my definitions and explanations the volunteer agrees to participate in this study.

_____ _____

Date Name of investigator who gave the information about the study

Signature: _____