The evaluation of diagnostic tests for sexually transmitted infections

Max A Chernesky PhD

Diagnostic tests should receive method- and use-effectiveness evaluations. Method-effectiveness evaluations determine sensitivity, specificity and predictive values for new tests. Use-effectiveness evaluations determine how practical or convenient a new test will be in a specific setting and may not be performed in a formal way in North American laboratories. To perform a clinical method evaluation of diagnostic tests, a good relationship between laboratory and clinical personnel is essential. Studies are usually conducted separately on populations of men and women, and should include sampling from different prevalence groups. Test performance comparisons may be made on a single specimen type or on more than one specimen from the same patient, which allows for the expansion of a reference standard and includes the ability of a particular assay, performed on a specimen type to diagnose an infected individual. The following components of the evaluation should be standardized and carefully followed: specimen identification; collection; transportation; processing; quality control; reading; proficiency testing; confirmatory testing; discordant analysis – sensitivity, specificity and predictive value calculations; and record keeping. Methods are available to determine whether sample results are true or false positives or negatives. Use-effectiveness evaluations might determine the stability or durability of supplies and equipment; the logistics of shipping, receiving and storing supplies; the clarity and completeness of test instructions; the time and effort required to process and read results; the subjectivity factors in interpretation and reporting; and the costs. These determinations are usually more apparent for commercial assays than for homemade tests.

Key Words: Diagnostic tests; Method effectiveness; Predictive value; Sensitivity; Specificity; Standardization; Use effectiveness

Sexually transmitted infections (STIs), including HIV, are increasing in prevalence worldwide. The World Health Organization estimated that 340 million new cases of STIs occurred in 1999 (1). Cases continue to increase in both developed and developing countries. During the past 20 years, advancements in diagnostic technologies have provided new, more sensitive tests for several sexually transmitted organisms. The development and use of newer, more sensitive assays has occurred because commercial companies have invested in their development. This is most apparent in the development of assays for the diagnosis of Chlamydia trachomatis infections.

Laboratories using these new tests recognize the importance of knowing how they perform in comparison with tests already in use, and how reliably these tests provide useful information concerning the infectious status of a patient for treatment and management purposes. This has led to a large number of publications describing clinical evaluations of tests. It has also forced us to recognize the value of new assay approval by governmental agencies or peer review.

Method-effectiveness evaluations are performed to establish the sensitivity and specificity of new tests. Because tests may perform differently according to infection prevalence, positive and negative predictive values should also be calculated. Use-effectiveness evaluations are performed to determine how practical or convenient a test is in a particular setting. Protocols should be developed to examine the stability or durability of supplies and equipment, the logistics of shipping, receiving and storing supplies, the clarity and completeness of test instructions, the time and effort to process and read the results, the objectivity and subjectivity factors in test
interpretation and reporting, and the total costs, including reagents, supplies and personnel.

Before initiating evaluations, investigators should gather as much information as possible on the following: general demographic characteristics of the study population; the prevalence (or if prevalence is not available, the incidence) of the STI in the study population; prevailing STI treatment practices and antibiotic recommendations; the clinics and other sources of patient enrollment; methods or tests normally available or used for diagnosing the STI; and finally, the laboratories (including personnel) or other sites where testing will be performed.

**METHODS**

**Sampling of populations**

Evaluations should be performed separately on men and women, on individuals with and without symptoms, and in populations of varying prevalence. Individuals should be excluded if they have used antimicrobials preceding the start of the study if such treatment may invalidate test results. Confidence intervals, probabilities and statistical power calculations should be considered. Several good statistical references, both print- (2,3) and computer software-based (Solo Statistical system, BMDP Statistical Software Inc, USA), are available.

**Specimens**

The simplest comparison of a new test with an established one may involve only one specimen type. Because it is known that certain types of tests perform better or worse on a specific specimen type, comparing new tests with more than one specimen type provides more useful data for diagnosing an infected patient (4).

For men, specimens for assay evaluation are usually urethral swabs, first-void urine (the first 20 mL of any void) or rectal swabs. For urethral collections, a swab with a narrow metal shaft should be inserted 2 cm to 4 cm, rotated gently and withdrawn. For women, cervical, urethral or rectal swabs and first-void urine samples may be assayed. Other samples might be vulval, vaginal or introital swabs. Specimen selection is determined by the STI and the type of assay being evaluated.

The number of specimens needed depends on the number of assays under evaluation and the need for reference standard comparison. Where possible, specimen numbers from individual patients should be kept to a minimum, especially in asymptomatic patients, and any effect of the order of collection of specimens on test performance should be monitored, recorded and evaluated.

**Testing**

Testing procedures include identification, collection, transportation, processing, quality control, reading and proficiency testing.

**Identification**: A system for labelling specimens with unique identifiers should be established to maintain a link between specimens and patient records. These identifiers may include a study site number or code, a reference laboratory number and a patient identification number.

**Collection**: Specimens should be collected carefully (as described above) using sterile techniques. Where possible, the same collector should be used for the entire study to reduce variability due to technique. Where several collectors are involved, training should be used to standardize the collection process, and records should be kept on collector differences. Most commercial kits provide swabs and instructions for collection.

**The evaluation of diagnostic tests for STIs**

![Figure 1](link) Discordant analysis algorithm for the determination of positives to be included in a reference standard

**Transportation**: Most commercial transportation tubes have been optimized to preserve the analyte for testing. Package inserts should be followed concerning transportation and storage. Non-commercial methods such as culture should use optimal cold chain transportation conditions (4°C), with the objective of culturing as soon as possible within 24 h.

**Processing**: Assays for detecting antigens such as enzyme immunoassay, direct immunofluorescence assay, nucleic acid hybridization and nucleic acid amplification (NAA) should be performed exactly as set out in the package insert of commercial kits or in protocols provided by the noncommercial provider. Specimens for culture should be processed according to standard published methods.

**Quality control**: Commercial assays should have positive and negative controls performed with each run. Periodically, appropriate quality controls should be set up with each assay, including weak and strong positives and at least one negative.

**Reading**: When various assays are being evaluated, they should always be read in a blinded fashion. Operators should be properly trained in reading results; operator differences, when observed, should be recorded. Standards of scoring positives and negatives should be set and, where possible, quantitative scoring of results should be used.

**Proficiency testing**: At the beginning, middle and end of each study, a proficiency panel from a reference laboratory should be used to assess the performance of each assay technology under evaluation.

**Repeat testing, test confirmation and discordant analysis**

Specimens showing contamination or toxicity in culture should be diluted 1:2 and 1:4 and reincultured. Specimens producing results at the cutoff or in an established 'equivocal zone' should be repeated twice, and the consensus of the three determinations should be used. Some technologies, such as enzyme immunoassay, have a tendency to provide false-positive results and require confirmatory testing of positives. This can be done by an antibody blocking test (5). Before assuming that specimens which are negative in the reference assay but positive in the new test are false positives, they should be processed through a discordant analysis algorithm as shown in Figure 1. The ideal reference standard would establish that a specimen is a true positive by culture or by a nonculture test that has been confirmed by a different assay measuring a different component of the organism.
Chernesky
to determine whether the confirmatory test is performing ade-
quately, a random sample of specimens found to be negative in
both the reference test and the new test should be assayed using a
confirmatory test. For NAA tests, an appropriate confirmatory
assay should measure a different analyte (eg, a different nucleic
acid segment) than that measured in the test under evaluation (6).
Theoretically, all of the samples found to be negative in the new
test and the comparator test may not be negative in a third, con-
firmatory assay because samples with low levels of analyte may be
intermittently positive due to a sampling phenomenon (7). From
a practical standpoint, the random selection of one-quarter to
to one-third of the double-negative specimens should be adequate to
test in a third assay. Statistically, if 200 specimens are selected
from a sample of 800 and no positives are seen in the confirmatory
test, then the false-positive rate is less than 0.17% (P≤0.05).
Specimens found to be positive in the reference test and nega-
tive in the new test can be examined for inhibitors that interfere
with the new test. For example, inhibitors of NAA may simply be
removed by a 1:2 to 1:4 dilution of the specimen. Alternatively,
the presence of inhibitors is confirmed if the discordant specimen
remains negative even after it is ‘spiked’ with test organisms (8).

Record keeping
Collected patient data should include an identification number,
age, sex, presence or absence of signs or symptoms, date of speci-
cum collection and the identification of the collector. In the
laboratory, the receipt date of the specimen, conditions of storage and
processing date should be recorded. Standard recording forms
should be developed for each study. Field and laboratory findings
are entered into a database (preferably computerized) as the
data are generated. This routine entry of data, daily or weekly,
should be without delay.

Data analysis
Discordant analysis allows construction of an expanded reference
standard and calculation of sensitivity, specificity and predictive
values by 2×2 table analysis (Figure 2). Figure 2 illustrates how to
calculate sensitivity, specificity and predictive values from discor-
dant analyses. Without an expansion of the reference standard, the
sensitivity of the new test is calculated as a/g, where a represents
the number of specimens positive in both the new test and the ref-
erence standard test, and g represents the total number of positives
by the reference standard test (a+b+c). Specimens positive in the new
test but negative in the standard are represented by b of the 2×2
table and require confirmatory testing to determine whether they
are ‘true’ or ‘false’ positives. All specimens included in b that are
confirmed as true positives are then added to those in a, creating a
new, larger number of positives represented as A in the expanded
reference standard. The false positives are then represented by B.
This reference standard expansion manoeuvre increases the sensi-
tivity (A/G) of the new test and allows a truer understanding of the
sensitivity of the reference standard. Similarly, without expanding the
reference standard, the specificity of the new test would be calculated as d/h,
where d is the number of specimens found to be negative according to both the new test and the refer-
ence standard, and h represents the total number of negatives by
the reference standard test (b+d). After discordant analysis and refer-
ence standard expansion, H becomes a smaller number as true pos-
divots are determined and the specificity (D/H) of the new test and allows a truer understanding of
the specificity (D/H) increases.

REFERENCES
1. Global prevalence and incidence of selected curable sexually
transmitted infections: Overview and estimates. World Health
2. Ryan BF, Joiner BL. The estimation of confirmed intervals for
binomial proportion. Mini tab handbook, 3rd edn. California:
1988;296:1210-1.
polymerase chain reaction and ligase chain reaction and testing of
multiple specimens may contribute to lower assay sensitivities for
diagnosing Chlamydia trachomatis infection in women. Mol Cell
specificity of the chlamydiazyme test for Chlamydia trachomatis
Confirmatory polymerase chain reaction testing for
Chlamydia trachomatis in first void urine from
asymptomatic and symptomatic men. J Clin Microbiol
1993;30:2241-5.
7. Smieja M, Mahony JB, Goldsmith CH, Chong S, Petrich A,
Chernesky M. Replicate PCR testing and probit analysis
for detection and quantitation of Chlamydia pneumoniae
from pregnant and nonpregnant women inhibitory to
amplification of Chlamydia trachomatis nucleic acid by PCR,
ligase chain reaction, and transcription-mediated amplification:
Identification of urinary substances associated with

Figure 2) Expansion of the reference standard: Effect on sensitivity,
pecificity, and positive and negative predictive values of new diagnostic
tests