# Evaluation of Nucleic Acid Amplification Tests in the Absence of a Perfect Gold-Standard Test

## *A Review of the Statistical and Epidemiologic Issues*

*Alula Hadgu,[*] Nandini Dendukuri,[‡] and Joergen Hilden[†]*

**Abstract:** During the past 10 years, medical diagnostic testing for sexually transmitted infections (STIs) has changed markedly as a result of the rapid expansion and marketing of nucleic acid amplification tests (NAATs). Among such new DNA/RNA-amplification techniques are the polymerase chain reaction (PCR), the ligase chain reaction (LCR), and the transcription-mediated amplification (TMA) tests. Regrettably, the test evaluation process undergone by these tests has not always been rigorous or scientifically sound. Here, we review the controversy surrounding the statistical evaluation of these NAATs. We also review some of the traditional and recent statistical methods developed to estimate test sensitivity and specificity parameters in the absence of reliable gold-standard tests. In particular, we review the traditional latent class modeling approach that requires the assumption of independence between diagnostic tests conditional on the true disease status, and the more recent procedures that relax the conditional independence assumption. Finally, we apply some of these statistical modeling techniques to real data to estimate the sensitivity and specificity of a NAAT for *Chlamydia trachomatis*. On the basis of the latent class modeling approach with a pessimistic prior for culture sensitivity, the NAAT specificity estimate was 97.6% and, on the basis of an optimistic prior, the specificity was 95.3%. Similarly, the sensitivity estimates ranged from 88.1% to 89.6%.

(*Epidemiology* 2005;16: 604–612)

Submitted 28 January 2004; final version accepted 10 May 2005.

From the *Division of STD Prevention, Centers for Disease Control and Prevention, Atlanta, Georgia; †the Department of Biostatistics, the University of Copenhagen, Copenhagen, Denmark; and ‡Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Canada.

***Editors' note:*** A commentary on this article appears on page 595.

Correspondence: Alula Hadgu, Centers for Disease Control and Prevention, 1600 Clifton Road, Atlanta, GA 30333. E-mail: ahadgu@cdc.gov.

Copyright © 2005 by Lippincott Williams & Wilkins

ISSN: 1044-3983/05/1605-0604
DOI: 10.1097/01.ede.0000173042.07579.17

Medical diagnostic testing is an important aspect of health care, and it accounts for a substantial amount of health care expenditure.[1] Several authors have concluded that diagnostic tests often are designed poorly and evaluated inadequately before they are marketed for routine clinical use.[2–5] In fact, most new diagnostic technologies have not been assessed adequately to determine whether their application improves public health.[6] Because they have entered clinical use without rigorous and sound evaluation, many diagnostic tests have proven unreliable and sometimes useless in subsequent studies.[7–11]

During the past 10 years, medical diagnostic testing for sexually transmitted infections (STIs) has changed markedly as a result of rapid expansion of recent technologies in general and nucleic acid amplification tests (NAATs) in particular. Among such new DNA/RNA-amplification technologies are the polymerase chain reaction (PCR), the ligase chain reaction (LCR), the strand-displacement amplification (SDA), and the transcription-mediated amplification (TMA) tests. These tests are now used widely to diagnose STDs. Unfortunately, as with their predecessors (eg, the enzyme immunoassay tests and the direct fluorescent antibody tests), these NAATs have not been evaluated rigorously before being marketed and disseminated.

NAATs are designed to amplify nucleic acid sequences that are specific for the organism; unlike culture tests, NAATs do not require viable organisms. NAATs differ in their amplification methods and in their target nucleic acid sequences. For example, the LCR, SDA, and PCR tests amplify a *Chlamydia trachomatis* DNA sequence in the cryptic plasmid that is found in almost all strains of *C. trachomatis*, whereas the TMA test amplifies a ribosomal RNA target (23S). The fundamental characteristic of NAATs is their ability to produce a positive signal from as little as a single copy of the target DNA or RNA, which is the reason why we hypothesize that NAATs have good sensitivity, possibly at the expense of specificity.

Unfortunately, the hypothesized high sensitivity of NAATs has given rise to a controversial estimation approach



Copyright © Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.

called discrepant analysis, which has been found to produce overly optimistic estimates of sensitivity and specificity. This review article, which explains the problem and some of the solutions that have been proposed so far, is organized as follows. First, we briefly review the concept of discrepant analysis, primarily in the context of *C. trachomatis* testing. We then review statistical methods developed to estimate test performance parameters, such as sensitivity and specificity of screening or diagnostic tests, in the absence of reliable gold-standard tests. Finally, we discuss the public health implications of optimistically-biased estimates of sensitivity and specificity.

In this review article, for illustrative purposes, we focus on estimation of sensitivity and specificity for *C. trachomatis*. However, one of our main objectives is to discuss the estimation of test performance parameters in the absence of a gold-standard test, which is of general interest in regard to any evaluation of assays for the detection of disease or assessment of exposure.

## WHAT IS DISCREPANT ANALYSIS?

Because many readers of EPIDEMIOLOGY may not be familiar with discrepant analysis and certainly not with all the nuances of recent articles and controversies, we summarize them here. Before we describe discrepant analysis, let us first define few terms. Two criteria commonly used for assessing the accuracy of a diagnostic tests are its sensitivity and specificity. Sensitivity is the probability that a diagnostic test will be positive given the true diagnosis is positive and specificity is the probability that the diagnostic test will yield a negative result, given the true diagnosis is negative. Estimation of sensitivity and specificity is a simple matter when the true diagnostic status can be determined without error, ie, when we have a gold-standard test.[12] Unfortunately, for some diseases, such as infectious diseases, a gold-standard test may not exist, whereas for other diseases it may be prohibitively expensive or invasive.

For many years the alloyed gold-standard test (the imperfect reference test) for infectious diseases testing has been the isolation of the organism in cell cultures. Cell culturing is commonly believed to have nearly 100% specificity but less optimal sensitivity. Thus, the true infection status of the individual members of the sample tested is not known with certainty because of the imperfect nature of cell culture, the historically accepted reference test. Ignoring the imperfect nature of culture, and estimating sensitivity and specificity of a new test by comparing its performance with cell culture, results in biased sensitivity and specificity estimates of the new test. It is in this context that discrepant analysis has been introduced in a naive attempt to correct the bias of the sensitivity and specificity of the new test by identifying the truly infected patients that cell culture testing misses.[3,13–15]

In discrepant analysis, the apparent false-positive samples (NAAT-positive and cell culture-negative) are subjected to additional testing usually via the use of ancillary tests, such as the direct fluorescent antibody tests or another DNA-amplification test using primers for another target (such as, in the case of *C. trachomatis*, the major outer membrane protein). If any one of these additional tests yields a positive result, then the original NAAT-positive result is considered to be a true positive; and the original culture-negative result is considered a false-negative result.[3] This step is called discrepant resolution.

For example, van Doornum et al[16] undertook a study in which urine specimens from 237 women were tested for *Chlamydia* using LCR and cell culture testing; the results are shown in Table 1. The estimates of sensitivity and specificity, assuming cell culture as an alloyed gold standard-test, are 86.7% (13/15) and 94.6% (210/222), respectively. The heart of discrepant analysis is the selective reclassification of some or all of the cases in which there is a discordance between the test under evaluation and the alloyed gold-standard test, that is, cell B, cell C or both. Thus, cases apparently false positive

**TABLE 1.** Comparison of LCR and Cell Culture Assays for *C. trachomatis* in Urine Collected From 237 Women Attending an STD Clinic (adopted from van Doornum et al[16])

| | Cell Culture (Cervix) | | | Discrepant Analysis by MOMP-LCR | |
|---|---|---|---|---|---|
| **Plasmid-LCR Results** | **Positive** | **Negative** | **Total** | **Positive** | **Negative** |
| Positive | 13 (cell A) | 12 (cell B) | 25 | 25 (13+12) | 0 (12−12) |
| Negative | 2 (cell C) | 210 (cell D) | 212 | 2 | 210 |
| Total | 15 | 222 | 237 | 27 | 210 |

Culture-based sensitivity of LCR = (13/15) = 86.7%, specificity = (210/222) = 94.6%.
Discrepant analysis-based estimates of sensitivity = (25/27) = 92.6% and specificity = (210/210) = 100%.
STD indicates sexually transmitted disease.

© 2005 Lippincott Williams & Wilkins

Copyright © Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.

or negative are given an opportunity to be converted, whereas concordant ones (cells A and D) are accepted as valid.

Table 1 also shows the distribution of the 237 women after discrepant analysis. All of the 12 individuals in cell B (the culture-negative and LCR-positive cell) were subjected to additional testing by the major outer membrane protein test; and they were all positive and were then moved to cell A. According to van Doornum et al, the adjusted estimates of sensitivity and specificity of LCR by discrepant analysis are now 92.6% (25/27) and 100.0% (210/210), respectively, and these are the estimates that van Doornum et al take as final. Most published studies that compare amplification tests against tissue culture and employ discrepant analysis eventually reclassify over 95% of cell B samples to cell A after additional testing.[15]

This method of estimating test sensitivity and specificity has been severely criticized as biased and unscientific.[3,13–15,17,18] A recent FDA Draft Guidelines[19] document states that "discrepant analysis does not solve the bias problem. It is just a more complicated wrong solution." The *Journal of Clinical Microbiology*, a journal that had previously published several articles based on discrepant analysis, finally invited a commentary on the issue. In the invited commentary McAdam, criticized discrepant analysis as biased;[20] in subsequent correspondence he referred to the approach as flawed[21] and advised that it be avoided.[22] Despite its conceptual and logical problems, discrepant analysis has become a standard method for estimating the sensitivity and specificity of many diagnostic tests.[18,23–54]

## OTHER PERTINENT ISSUES IN THE EVALUATION OF NAATs

### Detection of Infection State Versus Detection of DNA/RNA

As mentioned previously, the main advantage of NAATs is their ability to produce a positive signal from as little as a single copy of the target DNA or RNA. However, the detection of a single DNA/RNA copy may not be a sufficient condition for the diagnosis of a current infection. The detection of one *C. trachomatis* target RNA or the detection of one *M. tuberculosis* target DNA may not necessarily imply the presence of these infections in a clinical sense and, thus, may not constitute an indication for treatment.[15] This point is important in light of the fact that these tests are susceptible to laboratory and aerosol contaminations. It is also possible that these tests could be amplifying dead micro-organisms in situ or a related nonpathogenic organism. For example, the primers used by some NAATs for *N. gonorrhoeae* may cross-react with nongonococcal *Neisseria* species.[55–57] Each of these pitfalls leads to false-positive results, ie, reduced specificity.

Other biologic evidence raises questions over whether the loss of specificity of NAATs is minimal, as proposed by proponents of discrepant analysis.[30,58–61] Schllinger et al[62] showed that of 13 index persons who were NAAT-positive but culture-negative for *Chlamydia* (cell B type individuals), none of their sex partners was positive. In recent unpublished CDC data from the National Health and Nutrition Examination Survey, approximately 30% of individuals who were positive for gonorrhea by the LCR test stated that they never had sex.

Another related issue, which was discussed at the 1997 FDA Advisory Group meeting on discrepant resolution,[63] is the claim of "substantial equivalence" between, say, a NAAT and a culture assay. Substantial equivalence means that a new device has the same intended use as the predicate device and is as safe and effective when substituted for it. The purpose of a culture assay is to detect the presence of a viable organism, whereas, the purpose of NAATs as stated in their respective package inserts is to detect ribosomal RNA or DNA (which may or may not be viable). It does not appear logical to compare 2 tests with such distinct primary functions. This difference has ramifications on the applicability of the concept of substantial equivalence, which in turn has implications for product labeling.

### NAATs and Issues of Reproducibility

Reproducibility studies within and between laboratories form a mandatory step in the development of new diagnostic tests. In particular, if test-retest agreement is only moderate, sensitivity and specificity cannot both be high. Thus, reproducibility data may contradict claims of high sensitivity and specificity. In 1997, Hadgu[3] stated that the test performance indices of the LCR and PCR tests for both Chlamydia and gonorrhea were exaggerated. He postulated that these tests may even suffer from reproducibility problems, which was subsequently confirmed.[64–67] Peterson et al[66] retested 37 samples with either an equivocal result by PCR (19 samples) or samples with a discrepant result between culture, PCR, and an antigen detection method. After repeat testing of such samples they showed that 29 of the samples had a different interpretation and that all the interpretation change was due to PCR except for one culture sample.

Recently Castriciano et al[65] showed that both the PCR and LCR tests suffer from lack of reproducibility on repeat testing. Of 1004 urine specimens assayed by LCR, 120 samples (108 positive and 12 equivocal samples) were retested by the same test. Upon repeat testing, 13 of the positives and all of the equivocal samples were negative, resulting in a 21% change of interpretation of such samples. Similarly, a study by Gronowski et al[67] demonstrated that a significant reproducibility problem could occur during routine use of the LCR assay for *C. trachomatis* and *N. gonor-*


© 2005 Lippincott Williams & Wilkins

Copyright © Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.

*rhoeae*. Possible causes of these variable results for NAATs include false positive hybridization during the detection assay, nonspecific priming in the amplification phase, amplicon contamination, and the presence of inhibitors.[67]

## STATISTICAL TEST EVALUATION IN THE ABSENCE OF A GOLD STANDARD

Appropriate statistical methods for estimating the sensitivity and specificity of a new test in the absence of a gold standard were proposed long before the advent of discrepant analysis. We shall briefly describe some of these methods, and illustrate them using the data in Table 1. Although these methods do not suffer from the inconsistencies of discrepant analysis, they are not without problems. By the very nature of the task, these methods rely on judiciously chosen but empirically unverifiable assumptions about the unknowns of the situation, such as the nature of the relationship between the 2 tests.

### Method 1: Alloyed Gold Standard Test With Known Sensitivity and Specificity

A bias correction can be applied when the sensitivity and specificity of the alloyed gold standard test are known. Let $Se_r$ and $Sp_r$ denote the sensitivity and specificity of the alloyed gold standard test and let $Se_t$ and $Sp_t$ represent the sensitivity and specificity of the new test, with P denoting the true disease prevalence. This method requires us to assume that the tests are conditionally independent. If we assume the tests are conditionally independent, then we get the following probabilities for the cell frequencies shown in Table 1:

Prob (cell A) = the probability that the alloyed gold standard is positive and the new test is positive.

$$= PSe_r\, Se_t + (1 - P)\,(1 - Sp_r)\,(1 - Sp_t)$$

Prob (cell B) = the probability that the alloyed gold standard is negative and the new test is positive.

$$= P(1 - Se_r)\, Se_t + (1 - P)\, Sp_r\,(1 - Sp_t)$$

Prob (cell C) = the probability that the alloyed gold standard is positive and the new test is negative.

$$= PSe_r\,(1 - Se_t) + (1 - P)\,(1 - Sp_r)\, Sp_t$$

Prob (cell D) = the probability that the alloyed gold standard is negative and the new test is negative.

$$= P(1 - Se_r)\,(1 - Se_t) + (1 - P)\, Sp_r\, Sp_t$$

Because $Se_r$ and $Sp_r$ are known, $Se_t$ and $Sp_t$ can be determined by the following simple algebraic functions.[68]

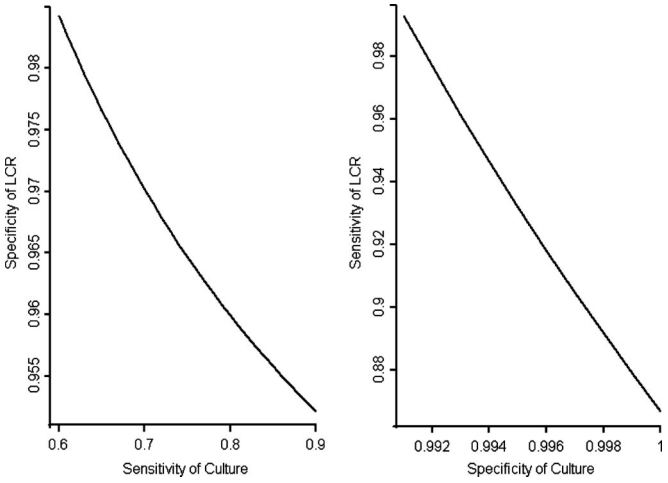


**FIGURE 1.** Variation in specificity and sensitivity of the LCR test with change in sensitivity and specificity of cell culture, respectively.

$$Se_t = \{Sp_r\,(A + B) - B\}/\{N(Sp_r - 1) + (A + C)\}$$

$$Sp_t = \{Se_r\,(C + D) - C\}/\{NSe_r - (A + C)\}$$

where N is the total sample size.

Figure 1 shows the variation in sensitivity and specificity of LCR over the plausible range of values of sensitivity and specificity of culture, for the data shown in Table 1. For example, if we knew for certain that the true sensitivity and specificity of culture are 75% and 100%, respectively, then the sensitivity of LCR is 86.7% and the specificity of LCR is 96.5%. Even though this method is easy to apply, its main limitation is that in most cases one does not know with certainty the true sensitivity and specificity of the alloyed gold-standard test.

### Method 2. Latent Class Analysis

Latent class analysis, which has been known for decades,[69] has recently found wide application for evaluating diagnostic tests in the absence of a gold-standard test. Examples include comparing the performance of diagnosticians in identifying dental caries,[70] evaluation of breast cancer screen tests,[71] colorectal cancer screening tests,[72] and diagnosis of *H. pylori*.[73,74] Latent class analysis is based on the concept that the observed results of different imperfect tests for the same disease are influenced by a common latent variable, the true disease status. Increasing the number of these tests increases our knowledge of the latent disease status, analogous to a large dark room becoming more illuminated with every additional light bulb turned on.

The traditional latent class analysis method is a generalization of the previous method (Method 1), which can be used in the absence of previous information on any of the

© 2005 Lippincott Williams & Wilkins

Copyright © Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.

parameters. This flexibility comes at the cost of requiring results from a minimum of 3 conditionally independent tests.[75] Thus it is not possible to apply this method to the data from van Doornum et al,[16] in which results on 3 tests were not available on all the subjects screened. Still, we present the theoretical framework of this model for our readers.

In the traditional latent class modeling approach, we assume that we have $p$ diagnostic tests for detecting a disease. For the $i^{th}$ diagnostic test, a positive test is denoted by $Y_i = 1$, and a negative result is denoted by $Y_i = 0$. Similarly, the true diagnosis is denoted by D = 0 or D = 1 for the absence or presence true disease, respectively. The probability of a positive response given D, is denoted by $\pi_{id} = \Pr(Y_1 = 1|D = d)$, $i = 1, \ldots p$, d = 0,1.

Note that, $\pi_{i1} = Se_i$ = sensitivity of the ith test and $\pi_{i0} = 1 - Sp_i = 1 -$ specificity. In addition let $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_p)$, represent the observed response vector. Then, $\Pr(\mathbf{Y}|D = d)$, d = 0,1, is the probability function of a multivariate Bernoulli distribution. Let $P_1$ denote the prevalence of disease, and $P_0 = 1 - P_1$. Then, the marginal probability of $\mathbf{Y}$ is given by

$$\Pr(\mathbf{Y}) = P_0 \Pr(\mathbf{Y}|d = 0) + P_1 \Pr(\mathbf{Y}|d = 1)$$

As mentioned previously, in traditional latent class models, it is assumed that the p diagnostic tests are conditionally independent of each other given the true disease status (the latent class D). If so, the unconditional probability of $\mathbf{Y}$ is given by

$$\Pr(\mathbf{Y}) = \sum_{d=0}^{1} P_d \prod_{i=1}^{p} \pi_{id}^{y_i} (1 - \pi_{id})^{1 - y_i}$$

Thus, the probability of the observed test profile is expressed as a function of the sensitivity, specificity, and prevalence parameters. The vector of unknown parameters $(\pi_{11}, \pi_{21} \ldots, \pi_{p1}, \pi_{10}, \pi_{20} \ldots, \pi_{p0}, P_0)$, is then obtained by the EM algorithm. Improvements of the traditional latent class analysis method have been developed in 2 directions as described below.

## Methods for Estimating Sensitivity and Specificity When Only 2 Tests Are Available

When only 2 tests (the new test and the alloyed gold standard) are available, we have too many unknown quantities that cannot be determined using the data alone. Statisticians call this a nonidentifiable problem. A statistical solution to such a problem can be obtained only through a Bayesian approach.[76] The Bayesian approach is becoming increasingly popular with medical researchers, particularly because it is easily interpretable and it provides a formal mechanism to combine information available prior to a study with the observed data.[77–79] Under the Bayesian framework, the marginal probability of the observed response vector $\mathbf{Y}$ remains the same as described in the case of the traditional latent class model. The unknown parameters are treated as random variables, each following a probability distribution. Information available on each parameter prior to collecting the data is summarized as the prior probability distribution, which is then combined with information from the observed data to obtain a posterior probability distribution for each parameter. The posterior distribution is essentially an update of our prior knowledge about each parameter, using the observed data. The posterior distribution can be used to obtain point estimates and credible intervals (ie, Bayesian confidence intervals) of sensitivity and specificity.

Prior distributions typically are determined from the published literature or in consultation with experts. To obtain a solution to this problem, prior information should be available on at least 2 parameters, typically the sensitivity and specificity of the alloyed gold standard tests. The form of this distribution usually is selected from probability distributions that lie in the range of 0 to 1, such as the uniform or beta distributions. In the case of cell culture, specificity is assumed to be close to 100% while a wide range of values have been reported for its sensitivity.[80–82] The results of the culture test are susceptible to quality control measures used while preserving and analyzing the sample. Therefore, we used 2 different prior distributions for the sensitivity of cell culture: (1) a pessimistic prior allowing equal weighting, (ie, a uniform distribution in the range 55%–65%), which would be reasonable in situations when we have low confidence in the implementation of quality control measures, and (2) an optimistic prior allowing equal weighting in the range 80% to 90%, which would be reasonable with proper quality control procedures.

A prior allowing equal weighting in the range 98% to 100% was used for the specificity in both cases. Prior distributions for sensitivity and specificity of LCR allowed equal weightage in the 0% to 100% range because it was assumed that nothing was known about the new test. The posterior distribution was obtained using a numerical algorithm called Gibbs sampling.[83] Figure 2 displays the posterior distributions of the sensitivity and specificity of LCR under the pessimistic and optimistic prior distributions for the data shown in Table 1. With the pessimistic prior for culture the median specificity of LCR is 97.6%, with a 95% credible interval ranging between 94% and 100%. However, when assuming the optimistic prior for sensitivity of culture we obtain a posterior median of 95.3%, with credible interval of 92% to 98%. The posterior distribution of sensitivity of LCR is not much affected since it depends primarily on the specificity of culture. (The posterior medians for sensitivity of LCR assuming the pessimistic and optimistic priors are 88.1% and 89.6%, respectively.)


© 2005 Lippincott Williams & Wilkins

Copyright © Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.

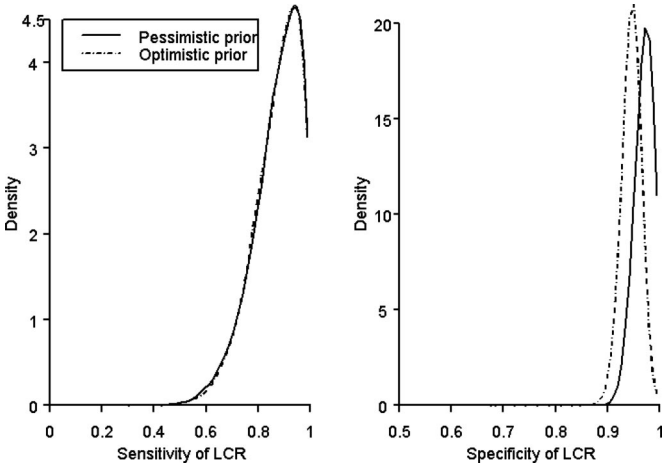


**FIGURE 2.** Posterior distributions of sensitivity and specificity of LCR under pessimistic and optimistic prior distributions of culture. For both parameters noninformative prior distributions were used.

Clearly, the results obtained with this method depend on the quality of the prior distribution used. The difference in our estimate of the specificity of LCR could have important implications regarding whether it would be considered a suitable screening test for an infection such as *Chlamydia*, which has a low prevalence in most populations. These results show that before a test is approved for widespread use, a range of plausible prior distributions need to be considered over all parameters. This approach has been recommended for Bayesian analysis of clinical trials.[84] If different prior opinions, held by various parties, lead to nearly the same final answers, the Bayesian methods can serve as a method of consensus formation.

### Methods for Modeling Conditional Dependence Between Tests

The traditional latent class model has been criticized for its assumption of conditional independence, which is difficult to justify in practice. When the conditional independence assumption is violated and there is a positive covariance between the tests, sensitivity and specificity can be substantially overestimated.[85] Recently, extensions to the traditional latent class models that relax the assumption of conditional independence have been proposed.[86,87] Furthermore, other researchers[80,88] propose models with the dual purpose of modeling conditional dependence while allowing for sensitivity and specificity to vary by covariates, since covariate-adjusted sensitivity and specificity can provide important information for determining target populations for screening. Increasing the number of parameters to be estimated in a latent class model increases the minimum number of tests required to obtain unique and reliable parameter estimates (model identifiability). This is a practical constraint in designing a study to evaluate the accuracy of a new test using such modeling approaches.

We now discuss one of the approaches[86,87] for modeling conditional dependence. In this modeling approach, Qu et al[86] add another latent variable R, which varies from subject to subject and has a standard normal distribution. This addition relaxes the conditional independence assumption by accounting for correlations among the tests. Thus, in latent class models with random effects the probability of a positive test is conditioned on both the latent class D and random effects R, and we use a probit model to describe the relationship:

$$\Pr(Y_i = 1 | D = d, R = r) = (\Phi a_{id} + b_{id} r),\ d=0, 1,\ R \sim N(0,1)$$

where $\Phi$ is the cumulative density function of the standard normal variate and $a_{id}$ and $b_{id}$ are unknown parameters. The probability of a positive test conditional on d only is:

$$\pi_{id} = \Pr(Y_i = 1 | D = d) = \int_{-\infty}^{\infty} \Phi(a_{id} + b_{id} r)\phi(r)dr = \Phi(a_{id}/(1 + b_{id}^2)^{1/2}).$$

Thus, the sensitivity and specificity for the $i^{th}$ test is given by:

$$Se_i = \pi_{i1} = \Phi(a_{i1}/(1 + b_{i1}^2)^{1/2}),$$
$$Sp_i = 1 - \pi_{i0} = \Phi(-a_{i0}/(1+b_{i0}^2)^{1/2})$$

In this approach a minimum of 4 diagnostic tests is required to obtain a unique solution. For model checking purposes, Qu et al[86] propose relevant statistics and graphical methods based on correlation residuals.

## IMPLICATIONS OF USING BIASED TEST PERFORMANCE INDICIES

Exaggerated indices of test performance have serious potential impact (Table 2). A false-positive test for STIs carries social, psychologic, and legal implications. In a 2001 device correction memo,[89] Abbott Labs stated that the specificity of their LCR test may be as low as 92% in some on-market lots and instructed their LCR Chlamydia assay users around the globe to retest all positive and equivocal samples. If the specificity of the LCR test is indeed as low as 92%, and if we assume the true prevalence is 4.0%, then as many as two-thirds of the women who tested positive would be false positives. Figure 3 shows the calculated positive predictive value (PPV) for a NAAT with 90% sensitivity and

© 2005 Lippincott Williams & Wilkins

Copyright © Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.

**TABLE 2.** Effect of Diagnostic Error and Biased Evaluation Practice for NAATs

| Effects of Untimely Adoption of and Reliance on a New Diagnostic Test (DT) Because of Optimistically Biased Evaluation Practice, as Exemplified by Discrepant Analysis | Effect of Relying on a (Conventional) Imperfect DT With Well-Characterized Performance as Exemplified by Cell Culture |
|---|---|
| Clinical management | Clinical management |
| 1. Unrecognized undertreatment or delayed treatment (ie, clinical sequella; contagion; epidemic; overly reassured patients) | Same problems |
| 2. Overtreatment (ie, increased cost; resistant micro-organisms; sensitization) | |
| Effect on public health | Effect on public health |
| 1. Biased incidence and prevalence estimates | Same problems but correction for bias possible |
| 2. Extent of undertreatment and overtreatment underrated | |
| 3. Degree of epidemic control overrated | |
| Effect on epidemiologic research | Effect on epidemiologic research |
| 1. Distorted results in studies of disease causation and disease propagation | Same problems but correction for bias possible |
| 2. Benefit/cost typically overestimated | |
| Innovation, marketing | Innovation, marketing |
| 1. Evaluation practice as an excuse for marketing practice | 1. Superiority of new and innovative DTs difficult to prove |
| 2. Marketing concerns may block proper evaluation | 2. Delayed acceptance of new and better DTs |
| 3. Good DTs may be ousted by inferior ones | 3. Unfair to serious industry and academic innovators alike |

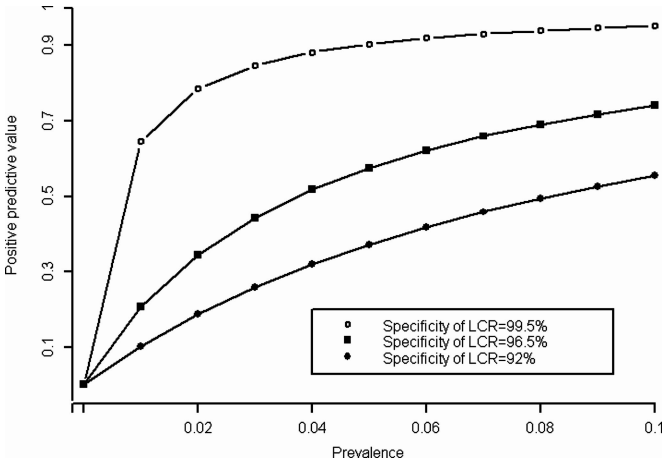


**FIGURE 3.** Positive predictive value for a NAAT with sensitivity of 90% and specificity of 92%–99.5%.

3 specificity estimates: 92% (from Abbott laboratory's device correction memo), 96.5% (based on latent class analysis in the previous section), and 99.5% (a discrepant analysis-based estimate which we believe is an overestimate). Note the steep decrease in PPV at lower prevalence levels, levels that characterize a substantial proportion of Chlamydia and gonorrhea screening programs in the United States.

Moreover, when the test is used in research, (be it therapeutic, prognostic, or epidemiologic), a small-but-unnoticed loss in specificity can greatly bias a prevalence estimate or an odds ratio. It can also reduce statistical power. From a clinical point of view, false positives can result in overtreatment. And, according to Hilden, "over-treatment, besides financial cost and unnecessary side effects, may cause allergic sensitization, which is catastrophic or immaterial depending on the patient's probability of subsequently needing the medicine."[14] Drugs used to treat *Chlamydia* rarely cause allergic reactions. In certain situations, however, false-positive NAAT results for say, *M. tuberclosis*, could delay treatment of other serious conditions such as cancers.[90] Additionally, it is important that in legal proceedings dealing with STIs and sexual abuse cases, the diagnostic tests must be highly specific.

## CONCLUSION

Insofar as the public health community is committed to keeping STIs in check and also to providing industry with a market for diagnostic tests, the present discussion boils down to 3 principal concerns. What are the consequences of using established diagnostic tests instead of adopting promising but not so well-characterized ones? What, on the other hand, are the consequences of adopting promising new tests on the basis of scientifically unjustifiable methods like discrepant analysis? Finally, how does one prove that the new test does indeed constitute an advance when there is no reliable gold standard?

Table 2 addresses the first 2 principal concerns and speaks for itself. Regarding the third concern, NAATs have shown superior sensitivity in the laboratory.[58] Quite understandably, this has led many researchers to the conviction that NAATs are preferable in clinical practice. However, it was


© 2005 Lippincott Williams & Wilkins
Copyright © Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.

difficult to demonstrate this superiority in clinical trials when there is no true gold standard. The recourse has been discrepant analysis. But now that discrepant analysis has been shown to be biased, other analytic strategies are needed. But now that discrepant analysis has been shown to be biased, other analytic strategies are needed.

Latent class models, and in particular their recent extensions[80,86–88] are useful tools. However, these methods have failed to account for the nature of the tests being evaluated. Hadgu and Dendukuri[91] proposed a hierarchical latent class model for evaluating NAATs that hypothesizes that tests based on different biologic phenomena are measuring different latent variables (eg, detection of DNA/RNA vs. the detection of current infection). This approach allows estimation of sensitivity and specificity with respect to each latent variable. However, all these sophisticated models are bound to make assumptions that may not be amenable to empirical verification. Therefore, we must navigate very carefully between over-reliance on laboratory data and insistence that field data should speak alone. This judicious weighing can be achieved only through a diligent and sustained interdisciplinary collaboration (versus consultation) between statisticians, clinicians, and laboratory scientists.

## REFERENCES


1. Begg CB. Biases in the assessment of diagnostic tests. *Stat Med*. 1987;6:411–423.
2. Reid CM, Mark LS, Feinstein AR. Use of methodological standards in diagnostic test research. *JAMA*. 1995;274:645–651.
3. Hadgu A. Bias in the evaluation of DNA-amplification tests for detecting *Chlamydia trachomatis. Stat Med*. 1997;16:1391–1399.
4. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *N Engl J Med*. 1978;299: 926–930.
5. Power EJ, Tunis SR, Wagner JL. Technology assessment and public health. *Annu Rev Public Health*. 1994;15:561–579.
6. Guyatt GH, Tugwell PX, Feeny DH, Haynes RB, Drummond M. A framework for clinical evaluation of diagnostic technologies. *CMAJ*. 1986;134:587–594.
7. Fletcher RH. Carcinoembryonic antigen. *Ann Intern Med*. 1986;104: 66–73.
8. Nierenberg AA, Feinstein AR. How to evaluate a diagnostic marker test. Lessons from the rise and fall of dexamethasone suppression test. *JAMA*. 1988;259:1699–1702.
9. Golightly MG. Laboratory considerations in the diagnosis and management of Lyme borreliosis. *Am J Clin Pathol*. 1993;99:168–174.
10. Lensing AW, Hirsh J. 125I-fibrinogen leg scanning: reassessment of its role for the diagnosis of venous thrombosis in post-operative patients. *Thromb Haemost*. 1993;69:2–7.
11. Rozanski A, Diamond GA, Berman D, Forrester JS, Morris D, Swan HJ. The declining specificity of exercise radionuclide ventriculography. *N Engl J Med*. 1983;309:518–522.
12. Pepe MS. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford: Oxford University Press; 2003.
13. Miller WC. Bias in discrepant analysis: when two wrongs don't make it a right. *J Clin Epidemiol*. 1998;51.
14. Hilden J. Discrepant analysis—or behavior? *Lancet*. 1997;350:902.
15. Hadgu A. Discrepant analysis. A biased and an unscientific method for estimating test sensitivity and specificity. *J Clin Epidemiol*. 1999;12: 1231–1237.
16. van Doornum GJ, Buimer M, Prins M, et al. Detection of Chlamydia trachomatis infection in urine samples from men and women by ligase chain reaction. *J Clin Microbiol*. 1995;33:2042–2047.
17. Hadgu A. The discrepancy in discrepant analysis. *Lancet*. 1996;348: 592–593.
18. Sternberg M. Discrepant analysis is still at large. *J Clin Microbiol*. 2001;39:826–827.
19. Food and Drug Administration. *Statistical Guidelines on Reporting Results From Studies Evaluating Diagnostic Tests*. Washington, DC: FDA; March 2003.
20. McAdam AJ. Discrepant analysis: how can we test a test? *J Clin Microbiol*. 2000;38:2027–2029.
21. McAdam AJ. Discrepant analysis is still at large [response]. *J Clin Microbiol*. 2001;39:826–827.
22. McAdam AJ. Discrepant analysis is an inappropriate and unscientific method [response]. *J Clin Microbiol*. 2000;38:4301–4302.
23. Bass CA, Jungkind DL, Silverman NS, et al. Clinical evaluation of a new polymerase chain reaction assay for detection of Chlamydia trachomatis in endocervical specimens. *J Clin Microbiol*. 1993;31:2648–2653.
24. Lee HH, Chernesky MA, Schachter J, et al. Diagnosis of *Chlamydia trachomatis* genitourinary infection in women by ligase chain reaction assay of urine. *Lancet*. 1995;345:213–216.
25. Loeffelholz MJ, Lewinski CA, Silver SR, et al. Detection of *Chlamydia trachomatis* in endocervical specimens by polymerase chain reaction. *J Clin Microbiol*. 1992;30:2847–2851.
26. de Barbeyrac B, Pellet I, Dutilh B, et al. Evaluation of the Amplicor *Chlamydia trachomatis* test versus culture in genital samples in various prevalence populations. *Genitourin Med*. 1994;70:162–166.
27. Bauwens JE, Clark AM, Loeffelholz MJ, et al. Diagnosis of *Chlamydia trachomatis* urethritis in men by polymerase chain reaction assay of first-catch urine. *J Clin Microbiol*. 1993;31:3013–3016.
28. Jaschek G, Gaydos CA, Welsh LE, et al. Direct detection of *Chlamydia trachomatis* in urine specimens from symptomatic and asymptomatic men by using a rapid polymerase chain reaction assay. *J Clin Microbiol*. 1993;31:1209–1212.
29. Wiesenfeld HC, Uhrin M, Dixon BW, et al. Diagnosis of male Chlamydia trachomatis urethritis by polymerase chain reaction. *Sex Transm Dis*. 1994;21:268–271.
30. Schachter J, Stamm WE, Quinn TC, et al. Ligase chain reaction to detect Chlamydia trachomatis infection of the cervix. *J Clin Microbiol*. 1994; 32:2540–2543.
31. Bassiri M, Hu HY, Domeika MA, et al. Detection of Chlamydia trachomatis in urine specimens from women by ligase chain reaction. *J Clin Microbiol*. 1995;33:898–900.
32. Ching S, Lee H, Hook EW 3rd, et al. Ligase chain reaction for detection of *Neisseria gonorrhoeae* in urogenital swabs. *J Clin Microbiol*. 1995; 33:3111–4.
33. Smith KR, Ching S, Lee H, et al. Evaluation of ligase chain reaction for use with urine for identification of *Neisseria gonorrhoeae* in females attending a sexually transmitted disease clinic. *J Clin Microbiol*. 1995; 33:455–457.
34. Pfyffer GE, Kissling P, Wirth R, et al. Direct detection of Mycobacterium tuberculosis complex in respiratory specimens by a target-amplified test system. *J Clin Microbiol*. 1994;32:918–23.
35. Vuorinen P, Miettinen A, Vuento R, et al. Direct detection of *Mycobacterium tuberculosis* complex in respiratory specimens by Gen-Probe Amplified Mycobacterium Tuberculosis Direct Test and Roche Amplicor Mycobacterium Tuberculosis Test. *J Clin Microbiol*. 1995;33:1856–1859.
36. Vlaspolder F, Singer P, Roggeveen C. Diagnostic value of an amplification method (Gen-Probe) compared with that of culture for diagnosis of tuberculosis. *J Clin Microbiol*. 1995;33:2699–2703.
37. Schue V, Green GA, Monteil H. Comparison of the ToxA test with cytotoxicity assay and culture for the detection of *Clostridium difficile*-associated diarrhoea disease. *J Med Microbiol*. 1994;41:316–318.
38. De Girolami PC, Hanff PA, Eichelberger K, et al. Multicenter evaluation of a new enzyme immunoassay for detection of *Clostridium difficile* enterotoxin A. *J Clin Microbiol*. 1992;30:1085–1088.
39. Crouch CF. Enzyme immunoassays for IgG and IgM antibodies to *Toxoplasma gondii* based on enhanced chemiluminescence. *J Clin Pathol*. 1995;48:652–657.
40. Pronovost AD, Rose SL, Pawlak JW, et al. Evaluation of a new immunodiagnostic assay for *Helicobacter pylori* antibody detection: correlation with histopathological and microbiological results. *J Clin Microbiol*. 1994;32:46–50.
41. Graham DY, Evans DJ Jr., Peacock J, et al. Comparison of rapid serological tests (FlexSure HP and QuickVue) with conventional ELISA


© 2005 Lippincott Williams & Wilkins


Copyright © Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.

for detection of *Helicobacter pylori* infection. *Am J Gastroenterol*. 1996;91:942–948.
42. Edelstein PH, Bryan RN, Enns RK, et al. Retrospective study of Gen-Probe rapid diagnostic system for detection of legionellae in frozen clinical respiratory tract samples. *J Clin Microbiol*. 1987;25: 1022–1026.
43. Knigge KM, Babb JL, Firca JR, et al. Enzyme immunoassay for the detection of group A streptococcal antigen. *J Clin Microbiol*. 1984;20: 735–741.
44. Roseff SD, Campos JM. Detection of cytomegalovirus antibodies in serum using the TranSTAT-CMV and CMV Scan assays. *Am J Clin Pathol*. 1993;99:539–541.
45. Zweygberg WB, Landqvist M, Hokeberg I, et al. Early detection of cytomegalovirus in cell culture by a new monoclonal antibody, CCH2. *J Virol Methods.* 1990;27:211–219.
46. LeBar WD, Resek CM, Crist AE Jr., et al. Comparison of a rapid latex agglutination assay and a fluorescent–antibody technique for the detection of herpes simplex antibody. *Diagn Microbiol Infect Dis*. 1988;11:21–24.
47. Dascal A, Chan-Thim J, Morahan M, et al. Diagnosis of herpes simplex virus infection in a clinical setting by a direct antigen detection enzyme immunoassay kit. *J Clin Microbiol*. 1989;27:700–704.
48. Cromien JL, Himmelreich CA, Glass RI, et al. Evaluation of new commercial enzyme immunoassay for rotavirus detection. *J Clin Microbiol*. 1987;25:2359–2362.
49. Sambourg M, Goudeau A, Courant C, et al. Direct appraisal of latex agglutination testing, a convenient alternative to enzyme immunoassay for the detection of rotavirus in childhood gastroenteritis, by comparison of two enzyme immunoassays and two latex tests. *J Clin Microbiol*. 1985;21:622–625.
50. Dennehy PH, Gauntlett DR. Evaluation of a new enzyme immunoassay (TESTPACK rotavirus) for the detection of rotavirus in fecal specimens. *Diagn Microbiol Infect Dis*. 1988;11:201–203.
51. Wester JP, Holtkamp M, Linnebank ER, et al. Non-invasive detection of deep venous thrombosis: ultrasonography versus duplex scanning. *Eur J Vasc Surg*. 1994;8:357–361.
52. Moncada J, Schachter J, Hook EW, et al. The effect of urine testing in evaluations of the sensitivity of the Gen-Probe APTIMA Combo 2 Assay on endocervical swabs for *Chlamydia trachomatis* and *Neisseria gonorrhoeae. STD*. 2004;31:273–277.
53. Rusch-Gerdes S, Richter E. Clinical evaluation of the semiautomated BDProbeTec ET System for the detection of *Mycobacterium tuberculosis* in respiratory and nonrespiratory specimens. *Diag Microbiol Infect Dis*. 2004;48:265–270.
54. Pai M, Flores LL, Pai N, et al. Diagnostic accuracy of nucleic acid amplification tests for tuberculous meningitis: a systematic review and meta-analyses. *Lancet Infect Dis*. 2003;633–643.
55. Martin DH, Cammarata C, Van Der Pol B, et al. Multicenter evaluation of AMPLICOR and automated COBAS AMPLICOR CT/NG tests for *Neisseria gonorrhoeae*. *J Clin Microbiol*. 2000;38:3544–359.
56. Van Der Pol B, Martin DH, Schachter J, et al. Enhancing the specificity of the COBAS AMPLICOR CT/NG test for *Neisseria gono*rrhoeae by retesting specimens with equivocal results. *J Clin Microbiol*. 2001;39:3092–8.
57. Crotchfelt KA, Welsh LE, DeBonville D, et al. Detection of *Neisseria gonorrhoeae* and *Chlamydia trachomatis* in genitourinary specimens from men and women by a coamplification PCR assay. *J Clin Microbiol*. 1997;35:1536–1540.
58. Schachter J, Stamm WE, Quinn TC. Discrepant analysis and screening for *Chlamydia trachomatis*. *Lancet*. 1998;351:217–218.
59. Schachter J, Stamm WE, Quinn TC. Discrepant analysis and screening for *Chlamydia trachomatis*. *Lancet*. 1996;348:1308–1309.
60. Chernesky M, Sellors J, Mahony J. Bias in the evaluation of DNA-amplification tests for detecting *Chlamydia trachomatis*. *Stat Med*. 1998;17:1064–1066.
61. Green TA, Black CM, Johnson RE. Evaluation of bias in diagnostic-test sensitivity and specificity estimates computed by discrepant analysis. *J Clin Microbiol*. 1998;36:375–381.
62. The 2004 National STD Prevention Conference Program and Abstract Book, March 8–11, 2004, Philadelphia, Pennsylvania.
63. The Food and Drug Administration. Available at: http://www.fda.gov.ohrms/dockets/ac/98/transcript/3387t1.pdf; accessed June 6, 2005.
64. Nordbo SA, Lund K, Skjeldestad FE. Retesting and follow-up of first-catch urines from men yield variable results with three *Chlamydia trachomatis* nucleic acid amplification tests. *APMIS*. 2000; 108:725–728.
65. Castriciano S, Luinstra K, Jang D, et al. Accuracy of results obtained by performing a second ligase chain reaction assay and PCR analysis on urine samples with positive or near- cutoff results in the LCx test for *Chlamydia trachomatis*. *J Clin Microbiol*. 2002;40:2632–2634.
66. Peterson EM, Darrow V, Blanding J, et al. Reproducibility problems with the AMPLICOR PCR Chlamydia trachomatis test. *J Clin Microbiol*. 1997;35:957–959.
67. Gronowski AM, Copper S, Baorto D, et al. Reproducibility problems with the Abbott laboratories LCx assay for Chlamydia trachomatis and Neisseria gonorrhoeae. *J Clin Microbiol*. 2000;38:2416–2418.
68. Staquet M, Rozencweig M, Lee YJ, et al. Methodology for the assessment of new dichotomous diagnostic tests. *J Chronic Dis*. 1981;34:599–610.
69. Lazarsfeld PF, Henry NW. *Latent Structure Analysis*, New York: Houghton-Miffin; 1968.
70. Espeland MA, Handelman SL. Using latent class models to characterize and assess relative error in discrete measurements. *Biometrics*. 1989;45: 587–599.
71. Goldberg JD, Wittes JT. The estimation of false negatives in medical screening. *Biometrics*. 1978;34:77–86.
72. Walter SD, Frommer DJ, Cook RJ. The estimation of sensitivity and specificity in colorectal cancer screening methods. *Cancer Detect Prev*. 1991;15:465–9.
73. Delaney BC, Holder RL, Allan TF, et al. A comparison of Bayesian and maximum likelihood methods to determine the performance of a point of care test for Helicobacter pylori in the office setting. *Med Decis Making*. 2003;23:21–30.
74. Christensen AH, Gjorup T, Hilden J, et al. Observer homogeneity in the histologic diagnosis of Helicobacter pylori. Latent class analysis, kappa coefficient, and repeat frequency. *Scand J Gastroenterol*. 1992;27:933–939.
75. Walter SD, Irwig LM. Estimation of test error rates, disease prevalence and relative risk from misclassified data: a review. *J Clin Epidemiol*. 1988;41:923–937.
76. Joseph L, Gyorkos TW, Coupal L. Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. *Am J Epidemiol*. 1995;141:263–72.
77. Brophy JM, Joseph L. Placing trials in context using Bayesian analysis. GUSTO revisited by Reverend Bayes. *JAMA*. 1995;273:871–875.
78. Goodman SN. Toward evidence-based medical statistics. 1: the P value fallacy. *Ann Intern Med*. 1999;130:995–1004.
79. Goodman SN. Toward evidence-based medical statistics. 2: the Bayes factor. *Ann Intern Med*. 1999;130:1005–1013.
80. Hadgu A, Qu Y. A biomedical application of latent class models with random effects. *Appl Stat*. 1998;47:603–616.
81. Black CM. Current methods of laboratory diagnosis of Chlamydia trachomatis infections. *Clin Microbiol Rev*. 1997;10:160–184.
82. Barnes RC. Laboratory diagnosis of human chlamydial infections. *Clin Microbiol Rev*. 1989;2:119–136.
83. Gelman A. *Bayesian Data Analysis*, London: Chapman & Hall; 1995.
84. Spiegelhalter DJ, Freedman LS, Parmar MK. Applying Bayesian ideas in drug development and clinical trials. *Stat Med*. 1993;12:1501–11; discussion 1513–1517.
85. Vacek PM. The effect of conditional dependence on the evaluation of diagnostic tests. *Biometrics*. 1985;41:959–52.
86. Qu Y, Tan M, Kutner MH. Random effects models in latent class analysis for evaluating accuracy of diagnostic tests. *Biometrics*. 1996; 52:797–810.
87. Dendukuri N, Joseph L. Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests. *Biometrics*. 2001; 57:158–67.
88. Qu Y, Hadgu A. A model for evaluating sensitivity and specificity for correlated diagnostic tests in efficacy studies with an imperfect reference test. *J Am Stat Assoc*. 1998;93:920–928.
89. Abbott Laboratories. Device Correction Memo, LCx *Chlamydia trachomatis.* February 2001.
90. Trinker M, Hofler G, Sill H. False-positive diagnosis of tuberculosis with PCR. *Lancet*. 1996;348:1388.
91. Hadgu A, Dendukuri N. Modeling conditional dependence between multiple diagnostic tests: a hierarchical latent class model. The 2004 XXIInd International Biometric Conference. 11–16 July 2004. Cairnes, Queensland, Australia.


© 2005 Lippincott Williams & Wilkins

Copyright © Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.