Evaluation of Screening Tests for Detecting *Chlamydia trachomatis*

Bias Associated With the Patient-infected-status Algorithm

Alula Hadgu,[a] *Nandini Dendukuri,*[b,c] *and Liangliang Wang*[d]

Abstract: In recent years, the evaluation of nucleic acid amplification tests (NAATs) for detecting *Chlamydia trachomatis* and *Neisseria gonorrhea* is based on a methodology called the patient-infected-status algorithm (PISA). In the simplest version of PISA, 4 test-specimen combinations (comparator tests) are used to define the gold standard. If a person shows a positive result by any 2 or more of these 4 comparator tests, the person is classified as infected; otherwise, the person is considered to be uninfected. A new test is then compared with this diagnostic algorithm. PISA-based sensitivity and specificity estimates of nucleic acid amplification tests have been published in the medical and microbiologic literature and have been included in FDA-approved package inserts of NAATs for detecting *C. trachomatis*. Using simulations, we compare 2 versions of the patient-infected-status algorithm with latent-class models and an imperfect gold standard. We show that the PISA can produce highly biased test-performance parameter estimates. In a series of simulated scenarios, none of the 95% confidence intervals for PISA-based estimates of sensitivity and prevalence contained the true values. In addition, the PISA-based estimates of sensitivity and specificity change markedly as the true prevalence changes. We recommend that PISA should not be used for estimating the sensitivity and specificity of tests.

(*Epidemiology* 2012;23: 72–82)

A judgment on the clinical utility of a new diagnostic test depends on the quality of the test evaluation study design and statistical methods, as well as on the ability to replicate the findings in actual or simulated experiments. Studies designed to assess diagnostic test performance indices such as sensitivity and specificity can be particularly susceptible to poor methodology.[1–3]

Medical devices are an integral part of health care, accounting for 25% of all ambulatory care expenditures in the United States.[1] Unfortunately, inadequate attention has been given to the statistical problems of evaluating diagnostic studies. In addition, the device approval process and regulation of medical devices are different from the evaluation and regulation of drugs.[2] Perhaps that is why some diagnostic tests, after hopeful introduction into medical practice, have proven to be less useful than expected.

In the mid-1990s, Hadgu[3–5] showed that the sensitivity and specificity estimates of chlamydia tests obtained from a widely used estimation approach called discrepant analysis were biased. Subsequently, discrepant analysis has been widely criticized[6–9] and its use as a parameter estimation approach has declined markedly. In its place, another estimation approach has been proposed, called the patient-infected-status algorithm (PISA).[10–14] In PISA, multiple tests are used to define the "gold-standard," and the sensitivity and specificity of a new test is then compared with this gold standard.

PISA-based estimates of sensitivity and specificity have been used by test-kit manufacturers in Food and Drug Administration (FDA)-cleared package inserts of nucleic acid amplification tests (NAATs) for detecting *Chlamydia trachomatis* and *Neisseria gonorrhea*. In addition, results and interpretations based on this algorithm have been published in medical and microbiologic journals.[10–14] We investigated whether the prevalence, sensitivity, and specificity estimates obtained by PISA are biased. We explored this in the context of *C. trachomatis* testing. We also compared the PISA-based estimates with other parameter estimation approaches such as latent class models and imperfect gold standards.

METHODS

Four evaluation approaches have been used to estimate the sensitivity and specificity of NAATs for detecting *C. trachomatis*: (1) tissue culture as a presumably perfect reference test; (2) PISA; (3) PISA2, another version of PISA; and (4) latent class models.

1. Tissue Culture as a Perfect Reference Test

Submitted 16 April 2011; accepted: 6 September 2011.

From the [a]Division of STD Prevention, National Centers for Disease Control and Prevention, Atlanta, GA; [b]Technology Assessment Unit, McGill University Health Center, Montreal, Quebec, Canada; [c]Department of Epidemiology and Biostatistics, McGill University, Montreal, Quebec, Canada; and [d]Department of Statistics, University of British Columbia, Vancouver, British Columbia, Canada.

The views expressed in this article are those of the authors and do not necessarily reflect the views or policies of the CDC or the US Public Health Service.

Supported by a Chercheur Boursier Junior 2 award from the Fonds de la Recherche en Sante' du Quebec. The authors reported no other financial interests related to this research.

Correspondence: Alula Hadgu, Division of STD Prevention, Centers for Disease Control, 1600 Clifton Road, Atlanta, GA 30333. E-mail: AHadgu@cdc.gov.

Copyright © 2011 by Lippincott Williams & Wilkins

ISSN: 1044-3983/12/2301-0072
DOI: 10.1097/EDE.0b013e31823b506b

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

For many years, the isolation of an organism in cell culture has been regarded as the gold standard test for estimating the sensitivity and specificity of new tests to detect infectious diseases. However, although culture is commonly believed to have nearly 100% specificity, its sensitivity is less optimal. Ignoring the limitations of tissue culture sensitivity and estimating performance parameters of a new test by comparing it with cell culture as the gold standard will produce biased sensitivity and specificity estimates.[15] Because of this potential for bias in culture tests, discrepant analysis was introduced as a potential solution.[15] Discrepant analysis was subsequently shown to produce biased estimates.[3–8] Because both discrepant analysis and tissue-culture-based estimates are problematic, some researchers have proposed the patient-infected-status algorithm as an appropriate method for evaluating tests in the absence of a perfect gold standard.[10–14]

2. PISA as a Reference Test

We first describe the PISA as presented by Schachter et al.[11] This general version of PISA is based on 2 tests (usually 2 NAATs in the case of chlamydia), each carried out on 2 specimen types (usually urine and endocervical specimens, in the case of chlamydia). This test-specimen combination results in 4 possible test results and 16 test profiles (Table 1). If a person shows positive results by 2 or more of these 4 test-specimen combinations (comparator tests), then the person is classified as infected; otherwise, the person is considered uninfected. Thus, as Table 1 shows, those with test profiles 1–11 are classified as infected, and the remaining 5 are classified as uninfected. The sensitivity of a new test according to PISA would be the probability that the new test is positive among the subjects classified as infected by PISA. Similarly, PISA-based specificity would be the probability that the new test is negative among the subjects classified as not infected by PISA.

Recognizing that classification by PISA is in fact not perfect, one can calculate the sensitivity and specificity of PISA itself. The sensitivity of the PISA is the probability that a person has one of the profiles 1–11, given that the person is truly infected (a latent variable). Similarly, the specificity of the PISA is given by the probability that a person has one of profiles 12–16, given that the person is truly not infected (as presented in the Appendix).

3. PISA Version 2 (PISA2) as a Reference Test

Several versions of PISA have used 2 or more NAATs to define the infected patient. Patient-infected-status algorithm version 2 (PISA2) is a slight modification used by the FDA to clear some chlamydia and gonorrhea assays. We present it here as an example of a more conservative PISA criterion for defining the infected person. In addition to the requirements of PISA, PISA2 requires that at least one specimen from each test must be positive. Thus, profiles 10 and 11 are read as positive with PISA, but negative with PISA2 (Table 1). All persons read as negative with PISA remain negative with PISA2. PISA2 was used to estimate the sensitivity and specificity of the FDA-approved package inserts for the APTIMA *C. trachomatis* assay, APTIMA *Neisseria gonorrhea* assay, and the APTIMA COMBO 2 assay.[16–18]

TABLE 1. Infection Status Based on the Patient-infected-status Algorithm (PISA) and PISA2

	Test 1		Test 2			
Profile	Specimen 1 Comparator1	Specimen 2 Comparator2	Specimen 1 Comparator3	Specimen 2 Comparator4	PISA	PISA2
1	+	+	+	+	Infected	Infected
2	+	+	+	−	Infected	Infected
3	+	+	−	+	Infected	Infected
4	+	−	+	+	Infected	Infected
5	+	−	+	−	Infected	Infected
6	+	−	−	+	Infected	Infected
7	−	+	+	+	Infected	Infected
8	−	+	+	−	Infected	Infected
9	−	+	−	+	Infected	Infected
10	+	+	−	−	Infected	Uninfected
11	−	−	+	+	Infected	Uninfected
12	+	−	−	−	Uninfected	Uninfected
13	−	+	−	−	Uninfected	Uninfected
14	−	−	+	−	Uninfected	Uninfected
15	−	−	−	+	Uninfected	Uninfected
16	−	−	−	−	Uninfected	Uninfected

© 2011 Lippincott Williams & Wilkins

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

4. Latent Class Models Assuming Conditional Independence

Researchers have used latent class models to estimate test performance parameters in various biomedical applications.[15,19–25] Like PISA, latent class models use multiple imperfect tests to create a gold standard. However, unlike the PISA approach, latent class models assign a weight to each test as determined by a statistical model. The statistical model assumes that the diagnostic tests under investigation are all imperfect measures of an unobserved, underlying (latent) variable—the true disease status.[20]

In the simplest latent class models, the observed test results from each person are assumed to be independent given the true disease status (the assumption of conditional independence). Papers by Hadgu et al[15] and Hui and Zhou[25] have provided a broad review of methods (including latent class models) to estimate test-performance parameters in the absence of a gold standard. Criticism of latent class models has centered around the difficulty in assessing their assumptions. In addition, the sensitivity and specificity of a new test can be substantially overestimated when conditional independence is falsely assumed.[26,27] Furthermore, most studies collect data on a small number of diagnostic tests, resulting in too few degrees of freedom to check the model assumptions. Bayesian estimation of these models using subjective knowledge of the sensitivity and specificity of the culture test are a potential solution.[15]

Simulations to Compare the 4 Estimation Approaches

To assess the performance of the estimation approaches described earlier in the text, we carried out a simulation study. We simulated datasets with results from 7 hypothetical tests—4 comparator (or standard) tests used to define PISA and PISA2, 2 new tests under evaluation, and one test with properties resembling those of an imperfect gold standard. Datasets were simulated using 3 sets of values for the parameters, spanning conditions of poor, moderate, and good performance of the comparator tests used in PISA and PISA2. In the absence of a gold standard test for *C. trachomatis*, the actual accuracy of any test is not known with certainty. Thus, our assumed values for our simulated datasets cover various possibilities that may be encountered in practice. We applied the 4 estimation approaches to each simulated dataset and compared the resulting estimates of test sensitivity and specificity, and infection prevalence to the true values.

Our simulations were constructed as follows. Sample size was 10,000 in all scenarios. We considered a sample size of 10,000 to be adequate to illustrate the performance of the 4 approaches. When sample size is large, an unbiased estimation approach should provide point estimates close to the true values and confidence/credible intervals that include the true values. Under each set of conditions, we randomly generated binary test results from 7 tests assuming known prevalence, sensitivity, and specificity values. We assumed true prevalence to be 5% or 2%. In most US screening sites, chlamydia prevalence is $<5\%$.[20] In the first 3 simulation scenarios, we assumed that the 7 tests are conditionally independent (ie, the results of any one test are not determined by the results of the remaining 6 tests within the diseased or nondiseased populations) (Tables 2–5), consistent with the assumption of a traditional latent class model. Thus, the probability of each combination of results on the 7 tests can be expressed in terms of the prevalence, sensitivity, and specificity of the individual tests. For example, if we assume S1, S2,…,S7 are the sensitivities of the 7 tests, and SP1, SP2,…,SP7 are the specificities, then the probability of observing a negative result on test 1 and positive results on the remaining tests is as follows:

$$\text{Prevalence} \times (1 - S1)\, S2\, S3\, S4\, S5\, S6\, S7 + (1\text{-Prevalence}) \times SP1\, (1 - SP2)\, (1 - SP3)\, (1 - SP4)\, (1 - SP5)\, (1 - SP6)\, (1 - SP7).$$

In the first simulated dataset, we assumed that a group of 4 tests (the comparator tests) and a test that we labeled as New Test1 have poor true sensitivity and specificity (60% sensitivity and 60% specificity). Thus, the New Test1 was defined such that it has properties that closely resemble the comparator tests. The sixth test (labeled New Test2) has moderate sensitivity and high specificity (75% and 97.5%, respectively), resembling the profile of several tests described for *C. trachomatis*. The seventh test is assumed to be a standard test with known sensitivity and specificity of 80% and 99%, respectively. We called this the imperfect gold standard because it has properties comparable with those of cell culture.

In the second simulated dataset, we assigned each of the 4 comparator tests and New Test1 moderate sensitivity and specificity (70% sensitivity and 75% specificity). In the third simulated dataset, we assigned the 4 comparator tests and New Test1 good sensitivity and specificity values (95% sensitivity and 95% specificity). The test performance values of New Test2 and the imperfect gold standard (the seventh test) remain the same in the second and third sets of simulations. Note that New Test2 has a performance that is superior to the comparator tests in the first and second simulations, but inferior to them in the third simulation. We repeated the third simulated scenario with the prevalence set to 2%. In the 3 simulation scenarios, data were generated using the statistical software R (version 2.7.1).

In the third simulation scenario, we made the comparator tests and New Test1 conditionally dependent, with a mild pairwise correlation in the nondiseased population of 0.2 and with 95% sensitivity and 95% specificity (Table 6). The conditional dependence was assumed to arise from a single random effect on the specificity of all 4 comparator tests and the New Test1. We used the function *simuData* as described in the *lcmr* package by Wang and Dendukuri to generate all

 © 2011 Lippincott Williams & Wilkins

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

TABLE 2. Sensitivity and Specificity Estimates and Their Associated 95% Confidence/Credible Intervals (CIs) Obtained by Using 4 Estimation Approaches

Test	True Value	PISA[a] Estimate (95% CI)	PISA2[b] Estimate (95% CI)	Imperf GS Estimate (95% CI)	LCM Estimate (95% CI)
			Sensitivity		
Comparator1	60.0			54.3 (49.4–58.6)	57.1 (52.3–61.9)
Comparator2	60.0			55.3 (50.9–59.7)	57.5 (52.6–62.6)
Comparator3	60.0			54.7 (50.3–59.0)	57.5 (52.6–62.6)
Comparator4	60.0			55.1 (50.7–59.4)	60.7 (55.6–65.5)
New Test1	60.0	41.3 (39.9–42.6)	41.7 (40.2–43.2)	55.1 (50.7–59.4)	58.1 (53.3–63.1)
New Test2	75.0	8.0 (7.3–8.8)	8.5 (7.7–9.4)	60.0 (55.6–64.2)	72.9 (65.4–80.6)
Imperfect GS	80.0	6.8 (6.2–7.5)	7.4 (6.6–8.2)	—	72.3 (63.1–81.2)
			Specificity		
Comparator1	60			59.7 (58.7–60.7)	60.0 (59.0–61.0)
Comparator2	60			59.5 (58.5–60.5)	59.7 (58.7–60.7)
Comparator3	60			60.1 (59.1–61.1)	60.3 (59.3–61.4)
Comparator4	60			60.5 (59.5–61.5)	60.9 (59.9–61.9)
New Test1	60	59.6 (58.1–61.0)	59.7 (58.4–61.0)	59.9 (58.9–60.8)	60.1 (59.1–61.2)
New Test2	97.5	95.7 (95.0–96.2)	95.3 (94.7–95.8)	96.4 (96.1–96.8)	97.6 (97.0–98.1)
Imperfect GS	99.0	97.3 (96.8–97.7)	96.9 (96.4–97.3)	—	99.0 (98.6–99.5)
Prevalence	5.0	53.7 (52.7–54.7)	42.5 (41.6–43.5)	4.9 (4.5–5.3)	5.1 (4.6–6.6)

Data were simulated by assuming conditional independence and the comparator tests have poor true sensitivity and specificity values (60% true sensitivity and 60% true specificity).

PISA2 indicates a version of the patient infected status algorithm used in some FDA-approved chlamydia and gonorrhea package inserts; Imperf GS, imperfect gold standard test; LCM, latent class model.

[a]Sensitivity of PISA = 82.1; specificity of PISA = 47.5.

[b]Sensitivity of PISA2 = 70.6; specificity of PISA2 = 59.0.

simulated datasets.[28] The 95% confidence intervals for the sensitivity, specificity, and prevalence estimates obtained using PISA, PISA2, and the imperfect reference standard were based on Wilson's method.[29] All the latent class models were estimated using a Bayesian approach implemented by the *lcmr* package.[28] We report posterior median estimates and 95% credible intervals.

Finally, to compare the 4 estimation approaches in a practical situation, we applied the methods to data derived from a study[19] of 4583 women in the northwestern United States who were tested by 3 immunoassay tests (Micro Track EIA [Syva-EIA], Chlamydiazyme [Abbott-EIA] and Pathfinder [Sanofi-EIA]) and 3 nonimmunoassay tests (Micro Track DFA [Syva-DFA], Pace2 [Gen-Probe], and chlamydia culture). The purpose of this study was to evaluate the sensitivity and specificity of these 6 non-NAAT chlamydia tests. We used panels of 2 immunoassay tests and 2 nonimmunoassay tests as the comparator tests to define infected women. These data had previously been analyzed using a latent class model that assumed conditional dependence.[19]

RESULTS

Tables 2–5 show results from the simulated datasets assuming conditional independence. Table 2 shows results from the first simulated dataset. True disease prevalence is 5%; however, prevalence estimates by the PISA and PISA2 are 54% and 43%, respectively. None of the PISA- or PISA2-based confidence intervals contain the true sensitivity and prevalence values, indicating significant bias. In contrast, the prevalence estimates obtained through the imperfect gold standard and the latent class model estimation were closer to the true values. The good performance of the latent class model is not entirely surprising, given that the data were generated assuming conditional independence. The sensitivity estimates of the 2 new tests and the imperfect gold standard based on the 2 PISAs are considerably biased. For example, the true sensitivity of the imperfect gold standard is 80%, whereas sensitivity estimates based on the 2 PISAs are both 7%. However, the specificity estimates based on PISA and PISA2 are close to the true values and even included in the 95% confidence interval in the case of New Test1. The sensitivity estimates of the new tests based on the imperfect standard are also biased but not as much as those based on the PISAs (Table 2). The estimates of sensitivity and specificity based on the latent class model are close to the true values.

Table 3 shows the results from the second simulated dataset. Again, the prevalence and sensitivity estimates based on the 2 PISAs are substantially biased. For example, the true sensitivity values of New Test1, New Test2, and the imperfect gold standard test are 70%, 75%, and 80%, respectively, whereas estimates based on the PISA are 32%, 13%, and

© 2011 Lippincott Williams & Wilkins

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

TABLE 3. Sensitivity and Specificity Estimates and Their Associated 95% Confidence/Credible Intervals (CIs) Obtained by Using 4 Estimation Approaches

Test	True Values	PISA[a] Estimate (95% CI)	PISA2[b] Estimate (95% CI)	Imperf GS Estimate (95% CI)	LCM Estimate (95% CI)
			Sensitivity		
Comparator1	70.0			59.6 (55.2–63.9)	67.9 (63.1–72.4)
Comparator2	70.0			58.8 (54.3–63.1)	68.3 (63.4–72.8)
Comparator3	70.0			59.8 (55.4–64.1)	69.5 (64.9–74.0)
Comparator4	70.0			59.2 (54.7–63.5)	70.4 (65.6–74.9)
New Test1	70.0	31.9 (30.3–33.7)	32.8 (30.9–34.8)	57.9 (53.5–62.3)	67.1 (62.4–71.5)
New Test2	75.0	13.4 (12.2–14.6)	14.8 (13.4–16.3)	59.2 (54.7–63.5)	74.6 (69.7–79.4)
Imperfect GS	80.0	12.9 (11.8–14.2)	14.6 (13.2–16.2)	—	77.9 (72.5–83.0)
			Specificity		
Comparator1	75.0			75.2 (74.3–76.1)	75.6 (74.8–76.5)
Comparator2	75.0			74.7 (73.8–75.5)	75.1 (74.2–76.0)
Comparator3	75.0			74.4 (73.5–75.2)	74.9 (74.0–75.8)
Comparator4	75.0			74.5 (73.7–75.4)	75.1 (74.3–76.0)
New Test1	75.0	74.0 (73.0–75.0)	73.8 (72.8–74.7)	73.8 (72.9–74.7)	74.3 (73.4–75.2)
New Test2	97.5	96.9 (96.5–97.3)	96.4 (96.0–97.0)	96.6 (96.2–97.0)	97.4 (97.1–97.7)
Imperfect GS	99.0	98.5 (98.2–98.7)	98.0 (97.6–98.3)	—	98.9 (98.6–99.1)
Prevalence	5.0	29.2 (28.3–30.1)	22.3 (21.5–23.2)	4.9 (4.4–5.3)	4.9 (4.4–5.4)

The data were simulated assuming conditional independence and that the comparator tests have moderate true sensitivity and specificity values (70% sensitivity and 75% specificity).
[a]Sensitivity of PISA = 91.6; specificity of PISA = 73.8.
[b]Sensitivity of PISA2 = 82.8; specificity of PISA2 = 80.8.

TABLE 4. Sensitivity and Specificity Estimates and Their Associated 95% Confidence/Credible Intervals (CIs) Obtained by Using 4 Estimation Approaches

Test	True Value	PISA[a] Estimate (95% CI)	PISA2[b] Estimate (95% CI)	Imperf GS Estimate (95% CI)	LCM Estimate (95% CI)
			Sensitivity		
Comparator1	95.0			75.1 (71.1–78.7)	94.6 (92.6–96.4)
Comparator2	95.0			75.3 (71.3–78.9)	93.6 (91.4–95.6)
Comparator3	95.0			76.5 (72.6–80.0)	94.6 (92.4–96.4)
Comparator4	95.0			76.5 (72.6–80.0)	95.8 (93.9–97.3)
New Test1	95.0	75.5 (72.0–78.7)	80.8 (77.4–83.8)	75.3 (71.3–78.9)	93.7 (91.5–95.7)
New Test2	75.0	62.3 (58.4–65.9)	67.2 (63.3–70.9)	60.4 (56.1–64.6)	77.2 (73.5–80.7)
Imperfect GS	80.0	62.1 (58.3–65.8)	66.7 (62.8–70.4)	—	77.7 (74.0–81.1)
			Specificity		
Comparator1	95.0			93.5 (93.0–94.0)	94.6 (94.2–95.0)
Comparator2	95.0			94.3 (93.8–94.7)	95.3 (94.9–95.7)
Comparator3	95.0			93.7 (93.2–94.2)	94.7 (94.3–95.1)
Comparator4	95.0			94.1 (93.6–94.5)	95.1 (94.7–95.6)
New Test1	95.0	94.7 (94.2–95.1)	94.7 (94.2–95.1)	93.7 (93.1–94.1)	94.7 (94.2–95.1)
New Test2	97.5	97.5 (97.2–97.8)	97.5 (97.2–97.8)	96.5 (96.2–96.9)	97.5 (97.1–97.8)
Imperfect GS	99.0	98.9 (98.7–99.1)	98.9 (98.7–99.1)	—	98.9 (98.7–99.1)
Prevalence	5.0	6.4 (5.9–6.9)	5.9 (5.4–6.4)	5.0 (4.6–5.4)	5.0 (4.6–5.5)

The data were simulated assuming conditional independence and that the comparator tests have good true sensitivity and specificity values (95% sensitivity and 95% specificity).
[a]Sensitivity of PISA = 99.9; specificity of PISA = 98.6.
[b]Sensitivity of PISA2 = 99.5; specificity of PISA2 = 99.0.

13%, respectively. Similarly, sensitivity estimates based on PISA2 are 33%, 15%, and 15%, respectively.

When the 4 comparator tests have good sensitivity (95%) and specificity (95%) values, the sensitivity estimates obtained by the PISAs for all noncomparator tests are biased despite the fact that the sensitivity and specificity of PISA and PISA2 are close to perfect for this scenario (Tables 4 and 5). None of the PISA-based 95% confidence intervals for the

© 2011 Lippincott Williams & Wilkins

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

TABLE 5. Sensitivity and Specificity Estimates and Their Associated 95% Confidence/Credible Intervals (CIs) Obtained by Using 4 Estimation Approaches

Test	True Value	PISA[a] Estimate (95% CI)	PISA2[b] Estimate (95% CI)	Imperf GS Estimate (95% CI)	LCM Estimate (95% CI)
			Sensitivity		
Comparator1	95.0			58.9 (53.0–64.5)	92.6 (88.8–95.8)
Comparator2	95.0			60.3 (54.5–65.8)	95.1 (91.7–97.6)
Comparator3	95.0			58.5 (52.7–64.1)	93.1 (89.4–96.1)
Comparator4	95.0			59.9 (54.1–65.5)	93.9 (90.2–96.7)
New Test1	95.0	59.2 (54.0–64.1)	68.0 (62.3–72.9)	60.3 (54.5–65.8)	95.1 (91.9–97.6)
New Test2	75.0	47.0 (41.9–52.2)	53.4 (47.8–58.9)	47.2 (41.4–53.0)	75.2 (69.4–80.7)
Imperfect GS	80.0	49.0 (43.9–54.2)	56.3 (50.7–61.7)	—	80.1 (74.5–85.2)
			Specificity		
Comparator1	95.0			94.3 (93.9–94.8)	94.7 (94.3–95.1)
Comparator2	95.0			94.9 (94.4–95.3)	95.2 (94.8–95.6)
Comparator3	95.0			94.4 (93.9–94.8)	94.8 (94.3–95.2)
Comparator4	95.0			94.7 (94.2–95.1)	95.0 (94.6–95.4)
New Test1	95.0	94.8 (94.4–95.2)	94.9 (94.4–95.3)	94.5 (94.0–94.9)	94.9 (94.4–95.3)
New Test2	97.5	97.5 (97.2–97.8)	97.5 (97.2–97.8)	97.2 (96.8–97.5)	97.5 (97.2–97.8)
Imperfect GS	99.0	98.9 (98.6–99.1)	98.9 (98.7–99.1)	—	98.9 (98.7–99.1)
Prevalence	2.0	3.6 (3.2–3.9)	3.1 (2.8–3.4)	2.8 (2.5–3.2)	2.2 (1.9–2.4)

The data were simulated assuming conditional independence and that the comparator tests have good true sensitivity and specificity values (95% sensitivity and 95% specificity). In this table, the true prevalence is 2%.
[a]Sensitivity of PISA = 99.9; specificity of PISA = 98.6.
[b]Sensitivity of PISA2 = 99.5; specificity of PISA2 = 99.0.

TABLE 6. Sensitivity and Specificity Estimates and Their Associated 95% Confidence/Credible Intervals (CIs) Obtained by Assuming the Comparator Tests and New Test1 Are Conditionally Dependent

Test	True Value	PISA[a] Estimate (95% CI)	PISA2[b] Estimate (95% CI)	Imperf GS Estimate (95% CI)	LCM Estimate (95% CI)
			Sensitivity		
Comparator1	95.0			59.9 (53.9–65.6)	93.9 (89.9–96.6)
Comparator2	95.0			60.6 (54.6–66.3)	95.1 (91.3–97.7)
Comparator3	95.0			58.5 (52.5–64.3)	92.9 (88.7–96.9)
Comparator4	95.0			60.6 (54.6–66.3)	93.7 (89.9–96.5)
New Test1	95.0	51.6 (47.7–55.5)	55.7 (51.5–59.9)	59.9 (53.9–65.6)	95.7 (92.0–98.0)
New Test2	75.0	26.7 (23.4–30.4)	31.0 (27.2–35.1)	47.2 (41.2–53.2)	76.4 (69.8–82.2)
Imperfect GS	80.0	27.4 (24.0–31.0)	32.0 (28.2–36.2)		82.3 (76.0–87.8)
			Specificity		
Comparator1	95.0			94.2 (93.7–94.7)	94.5 (94.1–95.0)
Comparator2	95.0			94.6 (94.1–95.0)	94.9 (94.4–95.3)
Comparator3	95.0			94.3 (93.9–94.8)	94.7 (94.2–95.1)
Comparator4	95.0			94.5 (94.0–95.0)	94.8 (94.3–95.2)
New Test1	95.0	96.2 (95.8–96.6)	95.9 (95.5–96.3)	94.6 (94.2–95.1)	94.9 (94.5–95.4)
New Test2	97.5	97.5 (97.2–97.8)	97.5 (97.2–97.8)	97.2 (96.8–97.5)	97.5 (97.1–97.8)
Imperfect GS	99.0	98.9 (98.6–99.1)	98.9 (98.6–99.1)		98.9 (98.6–99.1)
Prevalence	2.0	6.5 (6.0–6.9)	5.5 (5.1–6.0)	2.8 (2.5–3.2)	2.1 (1.8–2.4)

The data were simulated assuming conditional dependence and that the comparator tests have good true sensitivity and specificity values (95% sensitivity and 95% specificity).
[a]Sensitivity of PISA = 99.9; specificity of PISA = 99.4.
[b]Sensitivity of PISA2 = 99.5; specificity of PISA2 = 99.6.

sensitivity or prevalence estimates contain the true values. When the test performance parameters of the comparator tests are kept constant but the prevalence is changed (to 2%), the PISA-based and imperfect gold standard–based estimates of sensitivity change markedly (Tables 4 and 5), indicating that these estimates (like predictive-value positives and negatives)

© 2011 Lippincott Williams & Wilkins

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

are functions of disease prevalence. The latent class model, on the other hand, models disease prevalence and therefore is not susceptible to this bias.

Figures 1 and 2 compare the true sensitivity and true specificity of PISA with the true sensitivity and specificity of the comparator tests. When the comparator true test sensitivity is ≥ 0.23, the true sensitivity of PISA is greater than the true sensitivity of the comparator tests (Fig. 1). However, when the true comparator test specificity is <0.77, the true PISA specificity is less than the comparator test specificity (Fig. 2). Thus, even if both the true sensitivity and specificity of PISA are very high, the PISA-based estimates of sensitivity and prevalence are biased. For example, in Table 5, the true sensitivity and specificity of PISA are 99.9% and 98.6%, respectively; however, PISA-based estimates of sensitivity and prevalence are severely biased.

To investigate the impact of conditional dependence on parameter estimation, we repeated the third simulation scenario by assuming the 4 comparator tests and New Test1 are conditionally dependent in the nondiseased population (Table 6). Even though the sensitivity and specificity of the comparator tests are good (95%), the PISA- and PISA2-based estimates of sensitivity and prevalence are still biased (Table 6). In particular, the PISA-based specificity estimate for New Test1, which is a conditionally dependent test, was overestimated, and the true values were not included in their 95% confidence intervals. For New Test2 and the imperfect gold standard test—which are not conditionally dependent on PISA—the sensitivity was grossly underestimated, with more extreme bias than in the conditional independence case (Table 5).

Results From an Evaluation Study of *C. trachomatis* Tests

Table 7 shows estimates of sensitivity and specificity of 6 *C. trachomatis* tests obtained by applying the 4 estimation methods to data from a previously published paper.[19] Based on the 2 versions of the PISA, all 6 tests except the Abbott-EIA test have "good" to "excellent" sensitivity and specificity estimates for detecting *C. trachomatis*. This is not consistent with results of previously published work, including the 2002 CDC STD Laboratory Guidelines.[19,30,31] For example, based on PISA2, the sensitivity estimates of the Syva-DFA, Pace2, and culture tests were 87%, 89%, and 99%, respectively, and all specificity estimates were >99%. If manufacturers of these non-NAAT tests had evaluated their tests using PISA2, their tests would have been considered to have excellent sensitivity and specificity. However, in previous published work, the sensitivity estimates of these non-NAATs are lower.[30] The sensitivity estimates based on a latent class model (that adjusts for dependence), and even the culture-based estimates, are less than the estimates obtained by the 2 PISAs.[19]

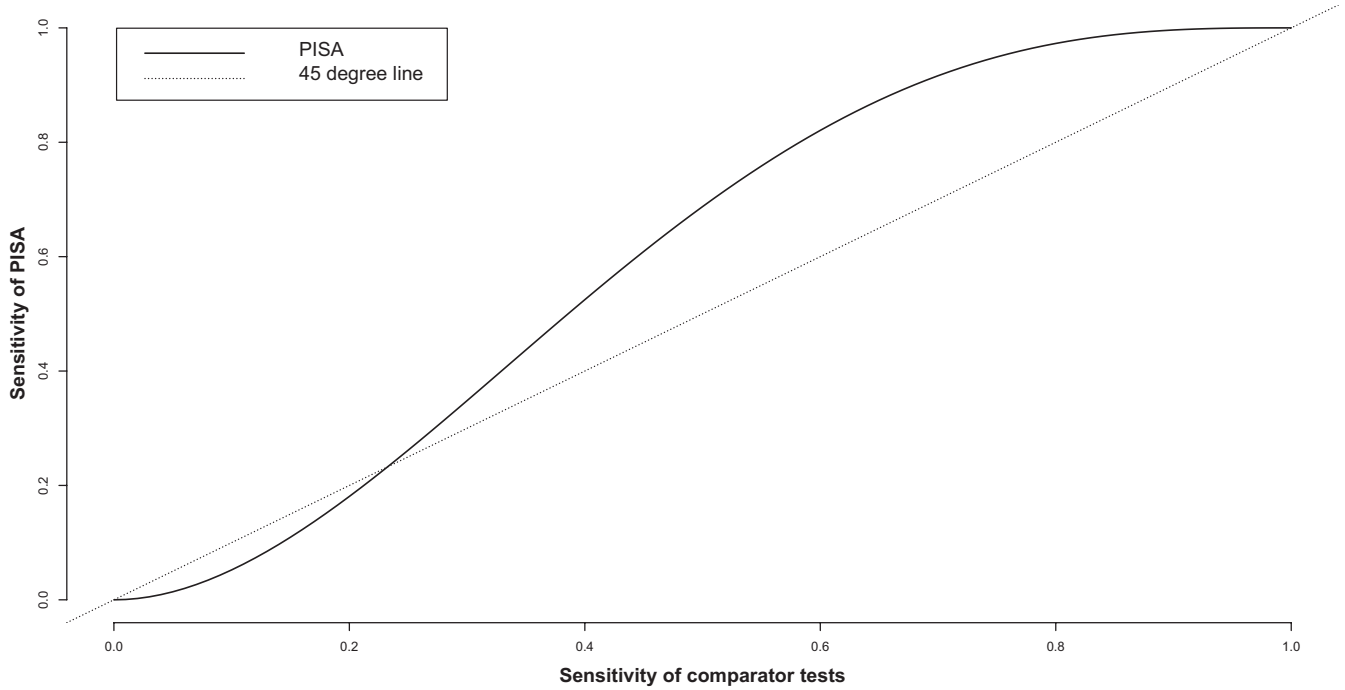

FIGURE 1. True sensitivity of PISA by true comparator test sensitivity.

© 2011 Lippincott Williams & Wilkins

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

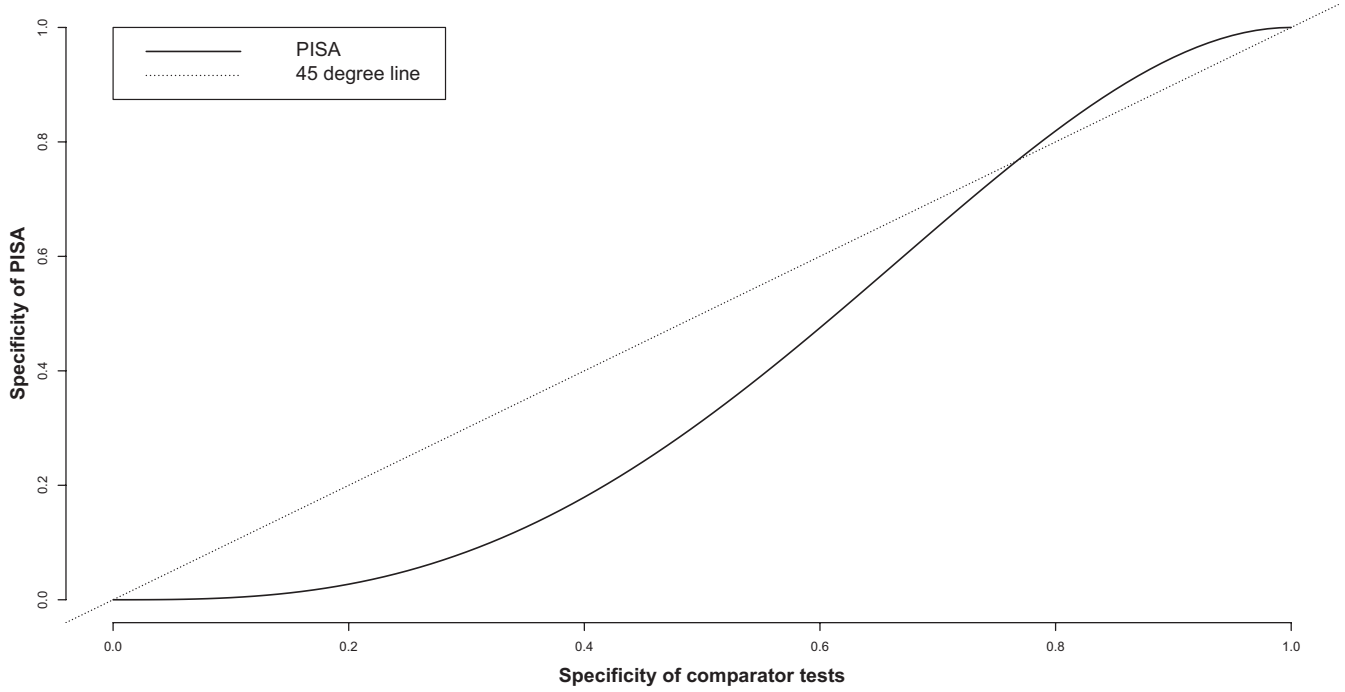

FIGURE 2. True specificity of PISA by true comparator test specificity.

TABLE 7. Sensitivity and Specificity Estimates of 6 *Chlamydia trachomatis* Tests Using Various Estimation Approaches[a]

Test	PISA Estimate (95% CI)	PISA2 Estimate (95% CI)	Culture Estimate (95% CI)	LCM Estimate (95% CI)
Sensitivity				
Syva–DFA	82.4 (75.4–87.9)	86.7 (79.8–91.6)	74.9 (67.9–80.8)	75.2 (68.8, 81.5)
Syva–EIA	82.4 (75.4–87.8)	83.9 (76.6–89.3)	73.3 (66.2–79.3)	73.8 (66.7, 80.3)
Abbott–EIA	71.7 (64.1–78.3)	74.2 (66.6–80.7)	63.6 (56.3–70.5)	65.7 (58.8, 72.1)
Gen–Probe	85.5 (79.1–90.3)	88.7 (82.5–93.0)	77.5 (70.8–83.2)	77.8 (71.0, 84.1)
Sanofi–EIA	77.2 (69.7–83.4)	80.1 (72.6–86.1)	67.9 (60.6–74.4)	69.1 (62.4, 75.8)
Culture	95.6 (90.7–98.1)	98.6 (94.6–99.8)	100	90.4 (83.9, 95.7)
Specificity				
Syva–DFA	99.6 (99.3–99.8)	99.4 (99.2–99.6)	99.8 (99.6–99.9)	99.8 (99.7, 99.9)
Syva–EIA	99.5 (99.2–99.7)	99.3 (99.0–99.5)	99.6 (99.4–99.8)	99.7 (99.5, 99.9)
Abbott–EIA	99.5 (99.3–99.7)	99.5 (99.2–99.7)	99.5 (99.3–99.7)	99.6 (99.4, 99.7)
Gen–Probe	99.5 (99.3–99.7)	99.5 (99.2–99.7)	99.6 (99.3–99.8)	99.6 (99.4, 99.8)
Sanofi–EIA	99.5 (99.2–99.7)	99.4 (99.1–99.6)	99.6 (99.4–99.8)	99.6 (99.4, 99.8)
Culture	99.2 (98.9–99.4)	99.0 (98.7–99.3)	100	99.7 (99.5, 99.9)

These data were first published in the *Journal of the Royal Statistical Society*.[16]

[a]For each test, 2 immunoassays and 2 nonimmunoassays were used as the 4 comparator tests for estimating the PISA and PISA2. The first 4 tests were used as the comparator tests to estimate the sensitivity and specificity of the last 2 tests. The last 4 tests were used as the comparator tests to estimate the sensitivity and specificity of the first 2 tests. Similarly, the first 2 and the last 2 tests were used to estimate the sensitivity and specificity of the 2 middle tests.

DISCUSSION

This study illustrates the bias due to an estimation approach called PISA, which is increasingly used to estimate sensitivity and specificity for tests of *C. trachomatis*.[10–14,16–18] The FDA has recommended against using the terms "sensitivity" and "specificity" to describe the comparison of a new test to a nonreference standard, and has recommended use of the terms "positive percent agreement" and "negative percent agreement."[32] However, recent package inserts continue to use the terms "sensitivity" and "specificity."

The PISA is a generalization of a method called the composite reference standard,[33] originally proposed to correct problems with discrepant analysis. However, as we demonstrated in our tables, such composite reference stan-

© 2011 Lippincott Williams & Wilkins

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

dards consisting of multiple imperfect reference tests can also result in biases.

One obvious problem with the patient-infected-status algorithm as applied to NAATs for detecting *C. trachomatis* is that the comparator tests and the new test to be evaluated are based on similar biologic and technologic processes; no attempt has been made to adjust for this dependence. For example, in both Schachter's[11] Table 2 and in the FDA-cleared package insert for the Gen-Probe's Aptima assay (Table 6d of the Aptima package insert),[16] the 4 comparator tests for a clinician-collected vaginal swabs were (1) Gen-Probe's Aptima Combo 2 Assay (based on cervical samples), (2) Gen-Probe's Aptima Combo 2 Assay (based on first-catch urine), (3) the BD Probe ET System (based on cervical samples), and (4) the BDProb ET System (based on first-catch urine). In other words, a "new" Aptima assay was evaluated by comparison with 2 existing Aptima assays with which it is likely to be correlated.

It is not reasonable to assume that NAATs using similar technologic methods and similar biologic specimens are conditionally independent. NAATs are known to have problems of false positivity resulting from stochastic contamination or systematic contamination in the laboratory.[15] These tests can also have problems of false negativity resulting from inhibitors.[12] These problems are likely to produce conditional dependence. If conditional dependence is ignored, test performance parameters can be substantially overestimated.[24,26,27]

Another shortcoming with the PISA is an associated implicit assumption that the detection of DNA by NAATs is equivalent to detection of current infection. This assumption is not consistent with the nature and intended use of NAATs.[34–36] NAATs detect a current or past marker of an organism (DNA or RNA) and not a disease state. A positive NAAT can indicate a current clinically active infection, the presence of residual DNA from a previous infection, the presence of residual DNA as a result of stochastic or systematic contamination in the laboratory, or a genuine false-positive result.[35]

A more realistic interpretation of NAATs is that the detection of chlamydial DNA or RNA is a necessary but not sufficient condition for current infection. Dendukuri et al[23] have proposed a novel statistical estimation method called the multiple latent variable model. Unlike previous methods, this approach allows estimation of sensitivity and specificity with respect to detection of DNA as well as the detection of current infection and adjusts for conditional dependence. However, although we can fit the chlamydia test evaluation data to a model that best represents our understanding of the problem, it is still difficult to prove that the proposed model is the correct model.[23]

PISA is an intuitively appealing procedure that, as our results show, can produce substantially biased estimates. Before estimating prevalence, sensitivity, and specificity for a new test, the methods and algorithms used must be demonstrated to be reliable and unbiased. Even if a reference test with true excellent sensitivity and specificity is used as an imperfect gold standard, the apparent sensitivity and the prevalence estimates of the new test can be substantially biased (Tables 4 and 5). This is because these methods do not account for prevalence in the parameter estimation process. Thus, despite the high specificity of PISA when the prevalence is low, the number of false positives increases markedly, affecting the apparent sensitivity estimates of the new test. In the prevalence levels we considered, the PISA-based specificity estimates were not so biased; however, the specificity estimates can be severely biased if prevalence is high.

Unlike positive and negative predictive values, sensitivity and specificity are inherent characteristics of a test and should not change with varying prevalence. However, the PISA-based sensitivity and prevalence estimates change markedly with changes in true prevalence (Tables 4 and 5). Thus, like earlier approaches (such as culture-based and discrepant analysis-based estimates of sensitivity and specificity), PISA-based estimates are biased. Absolute bias for sensitivity was as high as 72% (Table 2) when the sensitivity and specificity of the comparator tests were poor. Therefore, we recommend that PISA should not be used for estimating the sensitivity and specificity of new tests.

Appendix

Let T1, T2, T3, and T4 denote the 4 comparator tests. The true disease status is denoted by D. Each test, as well as the true disease status, can be positive (+) or negative (−). Following the definition of the patient infected status algorithm (PISA), if ≥2 tests in Table 1 are positive then the true infection status is positive. Let S1, S2, S3, and S4 denote the sensitivity of the 4 comparator tests. Similarly, let SP1, SP2, SP3, and SP4 denote the specificity of the 4 comparator tests. Finally, we denote probability by P.

If we assume that the tests that comprise the PISA are conditionally independent, then the sensitivity of PISA can be derived as follows:

$$\text{Sensitivity of PISA} = S_{PISA} = P(\text{profile1or profile2 or } \ldots \text{ profile11} \mid D+) = P(\text{profile1} \mid D+) + P(\text{profile2} \mid D+) + \ldots + P(\text{Profile11+} \mid D+), \text{ where}$$

$$\begin{aligned} P(\text{profile1} \mid D+) &= P(T1+, T2+, T3+, T4+ \mid D+) \\ &= P(T1+ \mid D+)\, P(T2+ \mid D+) \\ &\quad P(T3+ \mid D+)\, P(T4+ \mid D+) \\ &= S1\ S2\ S3\ S4 \\ P(\text{profile2+} \mid D+) &= P(T1+, T2+, T3+, T4- \mid D+) \\ &= P(T1+ \mid D+)\, P(T2+ \mid D+) \\ &\quad P(T3+ \mid D+)\, P(T4- \mid D+) \\ &= S1\ S2\ S3\ (1 - S4) \end{aligned}$$

© 2011 Lippincott Williams & Wilkins

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

$$\begin{aligned} P(\text{profile11}+ \mid D+) &= P(T1-, T2-, T3+, T4+ \mid D+) \\ &= P(T1- \mid D+)\, P(T2- \mid D+) \\ &\quad P(T3+ \mid D+)\, P(T4+ \mid D+) \\ &= (1 - S1)(1 - S2)\, S3\, S4 \end{aligned}$$

If we assume all the 4 comparator tests have the same sensitivity, S1, then,

$$S_{PISA} = (S1)^4 + 4(S1)^3 (1 - S1) + 6(S1)^2 (1 - S1)^2$$

Similarly,

$$\text{Sensitivity of PISA2} = S_{PISA2} = (S1)^4 + 4(S1)^3 (1 - S1) + 4(1 - S1)^2 (S1)^2$$

In the case of conditional dependence, the aforementioned expressions were aggregated across the random effects using Gaussian quadrature.[21]

If we assume that the tests that comprise the PISA are conditionally independent, then the specificity of the patient infected status algorithm (PISA) can be derived as follows:

$$\text{Specificity of PISA} = SP_{PISA} = P(\text{profile12 or profile13 or profile14 or profile15 or profile16} \mid D-) = P(\text{profile12} \mid D-) + P(\text{profile13} \mid D-) + \ldots + P(\text{profile16} \mid D-)$$

$$\begin{aligned} P(\text{profile12} \mid D-) &= P(T1+, T2-, T3-, T4- \mid D-) \\ &= P(T1+ \mid D-)\, P(T2- \mid D-) \\ &\quad P(T3- \mid D-)\, P(T4- \mid D-) \\ &= (1 - SP1)\, SP2\, SP3\, SP4 \\ P(\text{profile13} \mid D-) &= P(T1-, T2+, T3-, T4- \mid D-) \\ &= P(T1- \mid D-)\, P(T2+ \mid D-) \\ &\quad P(T3- \mid D-)\, P(T4- \mid D-) \\ &= SP1(1 - SP2)\, SP3\, SP4 \end{aligned}$$

…

$$\begin{aligned} P(\text{profile16} \mid D-) &= P(T1-, T2-, T3-, T4- \mid D-) \\ &= P(T1- \mid D-)\, P(T2- \mid D-) \\ &\quad P(T3- \mid D-)\, P(T4- \mid D-) \\ &= SP1\, SP2\, SP3\, SP4 \end{aligned}$$

If we assume all the 4 comparator tests have the same specificity, SP1, then,

$$SP_{PISA} = 4(SP1)^3(1 - SP1) + (SP1)^4$$

Similarly,

$$\text{Specificity of PISA2} = SP_{PISA2} = 4(SP1)^3(1 - SP1) + 2(SP1)^2 (1 - SP1)^2 + (SP1).^4$$

REFERENCES

1. Begg CB. Biases in the assessment of diagnostic tests. *Stat Med*. 1987;6:411–423.
2. Ramsey SD, Luce BR, Deyo R, Franklin G. The limited state of technology assessment for medical devices: facing the issues. *Am J Manage Care*. 1998;4(Spec no.):SP188–SP199.
3. Hadgu A. Bias in the evaluation of DNA-amplification tests for detecting *Chlamydia trachomatis*. *Stat Med*. 1997;16:1391–1399.
4. Hadgu A. The discrepancy in discrepant analysis. *Lancet*. 1996;348:592–593.
5. Hadgu A. Discrepant analysis: a biased and an unscientific method for estimating test sensitivity and specificity. *J Clin Epidemiol*. 1999;12:1231–1237.
6. Miller WC. Bias in discrepant analysis: when two wrongs don't make it a right. *J Clin Epidemiol*. 1998;51:1299–1303.
7. Hilden J. Discrepant analysis—or behavior? *Lancet*. 1997;350:902.
8. McAdam AJ. Discrepant analysis: how can we test a test? *J Clin Microbiol*. 2000;38:2027–29.
9. Food and Drug Administration. Guidance for industry and FDA staff: statistical guidance on reporting results from studies evaluating diagnostic tests. Bethesda (MD): US Health and Human Services; 2003. Available at: http://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm071287.pdf.
10. Martin DH, Nsuami M, Schachter J, et al. Use of multiple nucleic acid amplification tests to define the infected-patient "gold standard" in clinical trials of new diagnostic tests for *Chlamydia trachomatis* infections. *J Clin Microbiol*. 2004;42:4749–4758.
11. Schachter J, Chernesky MA, Willis DE, et al. Vaginal swabs are the specimens of choice when screening for Chlamydia trachomatis and Neisseria gonorrhoeae: results from a multicenter evaluation of the APTIMA assays for both infections. *Sex Transm Dis*. 2005;32:725–728.
12. Van Der Pol, Ferrero BD, Buck-Barrington L, et al. Multicenter evaluation of the BDProbeTec ET system for detection of *Chlamydia trachomatis* and *Neisseria gonorrhoeae* in urine specimens, female endocervical swabs, and male urethral swabs. *J Clin Microbiol*. 2001;39:1008–1016.
13. Gaydos CA, Quinn TC, Willis D, et al. Performance of the APTIMA Combo 2 assay for detection of *Chlamydia trachomatis* and *Neisseria gonorrhoeae* in female urine and endocervical swab specimens. *J Clin Microbiol*. 2003;41:304–309.
14. Chernesky MA, Martin DH, Hook EW, et al. Ability of new APTIMA CT and APTIMA GC assays to detect *Chlamydia trachomatis* and *Neisseria gonorrhoeae* in male urine and urethral swabs. *J Clin Microbiol*. 2002;43:127–131.
15. Hadgu A, Dendukuri N, Hilden J. The evaluation of nucleic acid amplification tests for detecting sexually transmitted diseases—a review of the statistical and epidemiological issues. *Epidemiology*. 2005;16:604–612.
16. GEN-PROBE APTIMA Assay for *Chlamydia trachomatis*. Available at: http://www.fda.gov/downloads/BiologicsBloodVaccines/SafetyAvailability/TissueSafety/ucm100235.pdf.
17. GEN-PROBE APTIMA Assay for *Neisseria gonorrhoeae*. Available at: http://www.gen-probe.com/pdfs/pi/501800RevA.pdf.
18. GEN-PROBE APTIMA Combo 2 Assay. Available at: http://www.gen-probe.com/pdfs/pi/501798RevA.pdf.
19. Hadgu A, Qu Y. A biomedical application of latent class models with random effects. *J R Stat Soc Ser C*. 1998;47:603–616.
20. Rindskop D, Rindskopf W. The value of latent class analysis in medical diagnosis. *Stat Med*. 1998;5:21–27.
21. Qu Y, Tan M, Kutner MH. Random effects models in latent class analysis for evaluating accuracy of diagnostic tests. *Biometrics*. 1996;52:797–810.
22. Dendukuri N, Joseph L. Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests. *Biometrics*. 2001;57:158–167.
23. Dendukuri M, Hadgu A, Wang L. Modeling conditional dependence between diagnostic tests: a multiple latent variable model. *Stat Med*. 2009;28:441–461.
24. Dendukuri N, Wang L, Hadgu A. Evaluating diagnostic tests for Chlamydia trachomatis in the absence of a gold-standard: a comparison of 3 statistical methods. *Stat Biopharm Res*. 2011;3:385–397.
25. Hui SL, Zhou XH. Evaluation of diagnostic tests without gold standards. *Stat Methods Med Res*. 1998;7:354–370.
26. Vacek PM. The effects of conditional dependence on the evaluation of diagnostic tests. *Biometrics*. 1985;41:959–968.
27. Torrance-Rynard VL, Walter SD. Effects of dependent errors in the assessment of diagnostic test performance. *Stat Med*. 1997;16:2157–2175.
28. Wang L, Dendukuri N. lcmr: An R package for Bayesian Estimation of Latent Class Models With Random Effects," Version 1.6.26. Available at: http://CRAN.R-project.org/package=lcmr Statistics.
29. Agresti A, Coull BA. Approximate is better than exact for interval estimation of binomial proportions. *Amer Statist*. 1998;52:119–126.
30. Newhall JW, Johnson RE, DeLisle et al. Head-to-head evaluation of five Chlamydia tests relative to a quality-assured culture standard. *J Clin Microbiol*. 1999;37:681–685.

© 2011 Lippincott Williams & Wilkins

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

31. Johnson RE, Newhall WJ, Papp JR, et al. Screening tests to detect *Chlamydia trachomatis* and *Neisseria gonorrhoeae* infections. *MMWR Recomm Rep*. 2002;51(RR-15):1–39.
32. FDA Statistical Guidance on Reporting Results from studies Evaluating Diagnostic Tests, 2007. Available at: http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm071148.htm?sms_ss==email#6.
33. Alonzo TA, Pepe MS. Using a combination of reference tests to assess the accuracy of a new diagnostic test. *Stat Med*. 1999;22:2987–3003.
34. Food and Drug Administration. Microbiology Devices Panel. Medical Advisory Committee Meeting. Rockville, MD: Miller Reporting Company; 2003. Available at: http://www.fda.gov/ohrms/dockets/ac/98/transcpt/3387t1.pdf.
35. Hadgu A. Issues in *Chlamydia trachomatis* testing by a nucleic acid amplification tests (Correspondence). *J Infect Dis*. 2006;193:1335–1336.
36. Persing DH, Tenover FC, Versalovic J, et al. Molecular Microbiology. Washington, DC: ASM Press; 2004.

 © 2011 Lippincott Williams & Wilkins

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