Head-to-Head Multicenter Comparison of DNA Probe and Nucleic Acid Amplification Tests for *Chlamydia trachomatis* Infection in Women Performed with an Improved Reference Standard

Carolyn M. Black,^{1*} Jeanne Marrazzo,² Robert E. Johnson,³ Edward W. Hook III,⁴
Robert B. Jones,⁵ Timothy A. Green,¹ Julius Schachter,⁶ Walter E. Stamm,²
Gail Bolan,⁷ Michael E. St. Louis,³ and David H. Martin⁸

National Center for Infectious Diseases¹ and National Center for HIV, STD, and TB Prevention,³ Centers for Disease Control and Prevention (CDC), Atlanta, Georgia; Department of Medicine, University of Washington, Seattle, Washington²; Department of Medicine, University of Alabama at Birmingham, Birmingham, Alabama⁴; School of Medicine, Indiana University, Indianapolis, Indiana⁵; Department of Laboratory Medicine, University of California,⁶ and San Francisco Department of Public Health,⁷ San Francisco, California; and School of Medicine, Louisiana State University Health Sciences Center, New Orleans, Louisiana⁸

Received 20 February 2002/Accepted 15 May 2002

Few evaluations of tests for *Chlamydia trachomatis* have compared nucleic acid amplification tests (NAATs) with diagnostic tests other than those by culture. In a five-city study of 3,551 women, we compared the results of commercial ligase chain reaction (LCR) and PCR tests performed on cervical swabs and urine with the results of PACE 2 tests performed on cervical swabs, using independent reference standards that included both cervical swabs and urethral swab-urine specimens. Using cervical culture as a standard, the sensitivities of PACE 2, LCR, and PCR tests with cervical specimens were 78.1, 96.9, and 89.9%, respectively, and the specificities were 99.3, 97.5, and 98.2%, respectively. Using either cervical swab or urine LCR-positive tests as the standard decreased sensitivities to 60.8% for PACE 2 and to 75.8 and 74.9% for PCR with cervical swabs and urine, respectively. Specificities increased to 99.7% for PACE 2 and to 99.7 and 99.4% for PCR with cervical swabs and urine, respectively. Sensitivities with a cervical swab-urine PCR standard were 61.9% for PACE 2 and 85.5 and 80.8% for LCR with cervical swabs and urine, respectively. Specificities were 99.6% for PACE 2 and 99.0 and 98.9% for LCR with cervical swabs and urine, respectively. Cervical swab versus urine differences were significant only for PCR specificities ($P = 0.034$). Overall, LCR sensitivity exceeded that of PCR, and sensitivities obtained with cervical swabs exceeded those obtained with urine specimens by small amounts. These data have substantiated, using a large multicenter sample and a patient standard, that LCR and PCR tests performed on endocervical swabs and urine are superior to PACE 2 tests for screening *C. trachomatis* infections in women. In our study, NAATs improved the detection of infected women by 17 to 38% compared to PACE 2.

Chlamydia trachomatis is the most prevalent bacterial sexually transmitted infection in the United States (6). Untreated infections can have devastating consequences for the female reproductive tract by causing pelvic inflammatory disease with serious sequelae such as infertility, ectopic pregnancy, and chronic pelvic pain. Because the majority of infections, particularly in women, are unrecognized clinically, screening with laboratory tests followed by antibiotic treatment is the most important means of intervention and control. The diagnostics industry has responded to this large public health need with steady improvements in test technologies. Concomitantly, methodology used for evaluation of these newly developed tests has evolved in a parallel fashion.

Historically, isolation of *C. trachomatis* in culture was used as the reference standard for classifying patients as infected or uninfected in new test evaluations (3). Culture was chosen for this purpose because it has nearly 100% specificity, resulting

from the highly distinctive morphology of specifically stained chlamydial inclusion bodies grown in host cells, and because it was the most sensitive method available at the time. Thus, virtually all of the early evaluations of commercial *C. trachomatis* tests were performed using culture alone as a reference standard. In most cases, these studies evaluated one new test against culture alone, a study design that made comparisons of new test performance among different laboratories difficult and that precluded evaluation of the performance of culture itself. Culture methods are technically difficult and not standardized; thus, performance can range widely (50 to 80% sensitivity), depending on the experience and skill of those in the laboratory (3). False-negative cultures can result in overestimation of the performance of an evaluated test if the test is more likely to detect culture-positive than culture-negative infections. Perhaps more importantly, false-negative cultures result in underestimation of the specificity of evaluated tests, an error which increases in magnitude as evaluated test sensitivity increases (25).

In an effort to overcome this problem and develop an improved reference standard, investigators in the mid-1980s began to consider additional test results when classifying the true

* Corresponding author. Mailing address: National Center for Infectious Diseases, Mailstop C17, Centers for Disease Control and Prevention, 1600 Clifton Rd. NE, Atlanta, GA 30333. Fax: (404) 639-3199. E-mail: cblack@cdc.gov.

infection status of subjects who tested as culture negative but tested as positive by a new test assay (2, 15, 18). Initially, other nonamplified tests were combined with culture as a reference standard (e.g., culture plus an antigen test such as direct fluorescent-antibody staining) (15, 18). This helped to improve the sensitivity of the reference standard, particularly in settings where cold-chain transport of specimens was problematic. However, no substantial improvement of the reference standard became available until the late 1980s, when the first nucleic acid amplification (NAA) technology (PCR) was developed. At this time, investigators combined one or more NAA tests (NAATs) with culture and sometimes also with direct fluorescent-antibody staining for use as a reference standard. This also addressed the conundrum posed by the enhanced sensitivity of the new NAA technology relative to that of culture. The increasing complexity of test combinations comprising the reference standard made clinical trials of new tests more impractical and expensive; consequently, most investigators applied the additional tests only to those patients whose specimens produced discrepant results for culture (negative result) and the new test (positive result). This design and the associated algorithms for classifying subjects as infected or uninfected are usually referred to as discrepant analysis. In 1996, the application of discrepant analysis to evaluation of tests for *C. trachomatis* was criticized as being statistically biased, because the results of the evaluated test influence the application of reference tests (12).

The choice of a reference standard is complicated by the fact that *C. trachomatis* can simultaneously infect multiple anatomic sites. A reference standard that includes both cervical swabs and urethral swab-urine specimens may better approximate the infection status of the patient (referred to as the patient standard) than a reference standard consisting of a test performed on only a single type of specimen (referred to as the specimen standard) (24).

We present here analyses utilizing a multitest reference standard that represents a patient standard and which avoids the use of discrepant analysis. We applied this reference standard to a head-to-head multicenter evaluation of commercial cervical swab and urine PCR and ligase chain reaction (LCR) tests and a widely used cervical swab DNA probe test (PACE 2) in a large and geographically diverse population of symptomatic and asymptomatic women attending sexually transmitted disease (STD) and family planning clinics. NAATs for *C. trachomatis* detection are generally considered to be more sensitive than nonculture, nonamplified tests, such as DNA probe or enzyme immunoassay, but the tests have been compared in few head-to-head studies. Our study objectives were (i) to compare LCR, PCR, and DNA probe tests for detection of *C. trachomatis* in cervical specimens; (ii) to compare LCR and PCR tests for detection of *C. trachomatis* in urine; (iii) to compare each NAAT for detection of *C. trachomatis* in cervical specimens with the same test for detection of *C. trachomatis* in urine; and (iv) to determine the effect on the above comparisons of modifying the reference standard by the following means: (a) adding urethral swab culture to a cervical swab culture reference standard and (b) substituting a cervical swab-urine NAAT reference standard for the culture standard. The last objective is both an assessment of the effect of replacing the specimen standard with a patient standard on estimates of

test performance and an evaluation of the use of NAATs to replace culture as a reference standard for test evaluation.

MATERIALS AND METHODS

Patient population. Female patients who presented to an STD clinic in each of the five participating centers (Birmingham, Ala., Indianapolis, Ind., New Orleans, La., San Francisco, Calif., and Seattle, Wash.) and a family planning clinic in New Orleans, La., with or without symptoms of an STD and with an indication for a pelvic examination, were eligible for enrollment. Pregnant women and women who had taken antibiotics in the previous 30 days were excluded. Study clinicians informed eligible patients of the study and enrolled those who consented to participate. The study was approved by institutional review boards at each research center and the Centers for Disease Control and Prevention. Enrollment began in October 1995 and ended in August 1997, when enrollment had reached the target sample of 407 culture-positive patients. If patients were enrolled more than once, only the results from the initial enrollment were included in this report. Enrolled patients were 63% African-American, 25% Caucasian, 4% Asian-Pacific Islander, and 7% other race or unknown. The median age was 24 years. Additional demographic and clinical data were collected and are reported elsewhere with analysis of risk factors for infection by type of test (J. Marrazzo, R. Johnson, W. Stamm, G. Bolan, E. Hook III, T. Green, J. Schacter, R. Jones, D. Martin, and C. Black, unpublished data). The study also included enrollment of males for whom results have been reported elsewhere (16).

Specimen collection. All tests described here detected *C. trachomatis* unless otherwise noted. Following the taking of the patient's history, patients were instructed to void without cleaning the perineal or periurethral area, saving the first 30 ml of urine in marked urine collection cups for LCR and PCR tests. Patients subsequently underwent a pelvic examination, during which a urethral swab sample was collected for culture by inserting the swab approximately 1 cm into the urethra with a rotary motion. Endocervical swabs were then collected by inserting a swab 2 to 3 cm into the cervical os and rotating the swab for 5 to 30 s. Endocervical swabs were collected in the following order: (i) a swab for Gram-stained smears and tests for *Neisseria gonorrhoeae*; (ii) randomly ordered swabs for DNA probe, LCR, and PCR; and (iii) a cytobrush specimen for culture. Endocervical swabs were transported in the test manufacturer's transport medium, except for swabs collected for PCR, which were transported in M4 transport medium (Multitest, Inc., Snellville, Ga.), and specimens collected for culture, which were transported as described below. Urine was held and transported to the laboratories at 4°C. LCR and PCR testing was completed within 4 days of specimen collection. In some cases for LCR testing, the specimens were frozen at -20°C or less and tested within 60 days.

Test methodology. Research laboratories at each center tested endocervical swab and urine specimens using the AMPLICOR microwell plate PCR test (Roche Diagnostic Systems, Branchburg, N.J.) and the LCx LCR test (Abbott Laboratories, Abbott Park, Ill.). Endocervical swabs and urine were processed, and the tests were performed according to the manufacturers' protocols, except that repeated LCR tests for subjects that had equivocal results on initial LCR testing were performed on specimens that had been frozen at -70°C for up to several weeks. A separate laboratory at each center tested endocervical swab specimens, using the PACE 2 DNA probe test (Gen-Probe, Inc., San Diego, Calif.).

Research centers used their own protocols for tissue culture isolation of *C. trachomatis*. Swabs were transported to the laboratory in sucrose phosphate chlamydia transport medium or Eagle's minimal essential medium in Earle salts, each containing fetal calf serum and antibiotics. Specimens were held at 4°C for a maximum of 18 to 72 h, depending upon the center, or frozen at -70°C. Culture medium was inoculated onto cycloheximide- and/or DEAE-dextran-treated McCoy cells in 96-well microtiter plates or 1-dram vials. Cultures were incubated at 35 or 37°C for 48 to 72 h. After incubation, cells were fixed with methanol or ethanol and stained with locally prepared or commercial major outer membrane protein-specific fluorescein-conjugated antibody reagents. One center used a lipopolysaccharide-specific antibody stain as well. Three of the centers performed a blind passage if no chlamydial inclusions were noted at 48 to 72 h.

Data analysis and reference standards. We compared estimates of sensitivity and specificity for each test using the following standards to classify patients as infected: (i) three different single cervical test reference standards (patients were classified as infected if the cervical culture, cervical LCR, or cervical PCR test result was positive) and (ii) three different two-test reference standards (patients were classified as infected if either the cervical or urethral culture, cervical or urine LCR, or cervical or urine PCR test result was positive). A two-test refer-

TABLE 1. Positivity of culture, LCR, PCR, and PACE 2 tests for *C. trachomatis* infection by research center and type of specimen

Type of specimen and test	Result for research center:											
	A (n = 564)		B (n = 743)		C (n = 884)		D (n = 771)		E (n = 589)		All (n = 3,551)	
	No. ^a	% ^b	No.	%	No.	%	No.	%	No.	%	No.	%
Cervical swab												
LCR	38	6.7	130	17.5	111	12.6	106	13.7	41	7.0	426	12.0
PCR	35	6.2	117	15.7	101	11.4	98	12.7	31	5.3	382	10.8
PACE 2	30	5.3	91	12.2	77	8.7	66	8.6	33	5.6	297	8.4
Culture	30	5.3	108	14.5	100	11.3	89	11.5	33	5.6	360	10.1
Urine												
LCR	34	6.0	131	17.6	86	9.7	117	15.2	38	6.5	406	11.4
PCR	38	6.7	126	17.0	92	10.4	77	10.0	41	7.0	374	10.5
Urethral swab												
Culture	17	3.0	61	8.2	67	7.6	60	7.8	15	2.5	220	6.2

^a Values are numbers of positive specimens.^b Values are percentages of the total (n).

ence standard in which subjects are classified as infected if either reference test is positive has been advocated to improve reference standard sensitivity and, thereby, reduce underestimation of evaluated test specificity due to false-negative reference standard results (1).

We consider the single-reference test to be a specimen standard, whereas the two-specimen standard can be regarded as a patient standard. Use of a patient standard may be particularly important to avoid bias when comparing the performance of cervical and urine tests. We compared the use of patient and specimen standards for comparing cervical tests with each other and urine tests with each other and for comparing cervical tests with urine tests. Since NAATs are approved for use with urine, we compared cervical with urine NAATs using the alternate technology NAAT and both the specimen and the patient standard, e.g., we compared cervical with urine LCR using the cervical PCR standard or the cervical or urine PCR standard. Lastly, because culture tests are technically difficult, costly, and not standardized, we assessed the effect on estimates of test performance of substituting a NAAT for culture in either a single-test specimen standard or a two-test patient standard.

Statistical methods. For each of the reference standards employed, sensitivity and specificity were calculated according to standard formulae (10). The statistical significance of center-to-center variation in sensitivity and specificity estimates was assessed using the Pearson chi-square test of homogeneity (21). We computed summary estimates of sensitivity and specificity, and their standard errors, by weighting the center-specific estimates, using the random effects model for combining proportions described by Laird and Mosteller (17). To evaluate center-to-center variation in the difference between sensitivity and specificity estimates for pairs of tests, we assigned each concordant specimen a score of zero and each discordant specimen a score of ± 1 , depending on which of the two test results was positive (negative). For each center, the mean of these scores is equal to the difference in sensitivity (specificity) estimates. The statistical significance of the variation in mean scores among centers was then assessed, using the analysis of variance *F* test (21). Summary estimates of these differences, as well as their standard errors, were computed using the weights derived for summaries of the component estimates.

RESULTS

Enrollment and number (percent) of subjects that tested positive. The number of women enrolled at each center ranged from 564 to 884, with a total of 3,551 (Table 1). The number of subjects testing as positive by cervical culture by center ranged from 30 to 108, with a total of 360. The percentage of cervical culture-positive subjects ranged by center from 5.3 to 14.5%. Urethral culture results were less frequently positive than those of all other tests ($P < 0.001$); the number of subjects that tested as positive by urethral culture ranged from 15 (2.5%) to 67 (7.6%) by center, with a total of 220 (6.2%).

Overall, the percentage of subjects positive by cervical tests was 10.1% by culture, 8.4% by PACE 2, 12.0% by LCR, and 10.8% by PCR. Overall, the number of LCR- and PCR-positive tests exceeded the number of PACE 2-positive tests by 43.4 and 28.6%, respectively ($P < 0.001$). The smallest center-specific increase in test results positive by LCR relative to those positive by PACE 2 was 24.2%, whereas the number of PCR-positive results for women for one center was lower than that for PACE 2. Overall, the number of culture-positive tests exceeded the number of PACE 2-positive tests by 21.2% ($P < 0.001$), but the percent positive was the same for culture and PACE 2 at two centers.

For cervical specimens, the percentage of LCR-positive tests significantly exceeded the percentage of PCR-positive tests at each center, with an overall increase of 11.5% ($P < 0.001$). For urine specimens, the percentage of LCR-positive tests exceeded the percentage of PCR-positive tests by 52% at one center ($P < 0.001$); at the remaining centers, the differences were inconsistent and none were significant. Overall, the percentage of positive tests by LCR or PCR with cervical specimens exceeded the percentage of positive tests with urine specimens by a small amount (0.6 and 0.2 percentage points, respectively), but the center-specific differences were not consistent with regard to which test produced more positive results or in the magnitude of the difference.

Head-to-head comparison of DNA probe and NAATs performed on cervical specimens. A head-to-head comparison of the sensitivities (Table 2) and specificities (Table 3) of the DNA probe, PCR, and LCR tests performed on cervical specimens was constructed using the independent patient standard of cervical culture plus urethral culture. Patients were classified as infected if either cervical or urethral culture result was positive. The sensitivities of the cervical NAATs significantly exceeded the sensitivity of PACE 2; cervical LCR sensitivity exceeded that of PACE 2 by 19.7 percentage points (95% confidence interval [CI], 12.9 to 26.6) and cervical PCR sensitivity exceeded that of PACE 2 by 12.4 percentage points (95% CI, 2.1 to 22.7) (Table 2). These differences represent proportional increases in sensitivity over PACE 2 of 28% for LCR and 17% for PCR. Conversely, the specificity of cervical

TABLE 2. Estimates of LCR, PCR, and PACE 2 sensitivities for detection of *C. trachomatis* infection by type of specimen and reference standard

Standard, test, and specimen type	No. of subjects infected	Sensitivity with ^a :				
		Cervical swab			Urine	
		LCR	PCR	PACE 2	LCR	PCR
One-specimen						
Culture, cervical swab	360	96.9 (94.2–99.6)	89.9 (84.3–95.6)	78.1 (70.1–86.2)	82.6 (75.6–89.6) ^b	79.2 (69.3–89.2) ^b
LCR, cervical swab	426		85.7 (81.5–89.8)	68.4 (60.9–76.0)		75.7 (67.3–84.2) ^b
PCR, cervical swab	382	95.8 (93.9–97.7)		74.9 (64.7–85.1)	83.6 (75.5–91.7) ^b	
Two-specimen patient (either specimen positive)						
Culture, cervical or urethral swab	407	91.4 (86.0–96.7) ^b	84.0 (80.5–87.6)	71.6 (62.0–81.3) ^b	83.4 (76.7–90.1) ^b	79.5 (69.2–89.7) ^b
LCR, cervical swab or urine	493		75.8 (71.8–79.9)	60.8 (51.4–70.2) ^b		74.9 (63.3–86.4) ^b
PCR, cervical swab or urine	462	85.5 (82.3–88.7)		61.9 (56.5–67.4)	80.8 (73.1–88.5) ^b	

^a Values are percentages. Values in parentheses are 95% confidence intervals.

^b *P* value for heterogeneity among research centers, <0.05. The estimates and 95% confidence intervals were derived from a random effects model (see Materials Methods).

PACE 2 significantly exceeded LCR specificity (difference = 1.3 percentage points [95% CI, 0.3 to 2.4] and PCR specificity (difference = 0.7 percentage points [95% CI, 0.1 to 1.3]) (Table 3).

Head-to-head comparison of LCR with PCR tests. A head-to-head comparison of PCR and LCR tests, as performed on cervical and urine specimens, with respect to their sensitivities (Table 2) and specificities (Table 3) was also constructed using the independent patient standard of cervical culture plus urethral culture. Although the sensitivities of cervical LCR tests exceeded those of cervical PCR tests at all centers and the sensitivities of urine LCR tests exceeded those of urine PCR tests at four of five centers, neither the difference for cervical specimen tests (difference = 7.3 percentage points [95% CI, 0.3 to 14.9]) nor that for urine tests (difference = 3.9 percentage points [95% CI, –4.7 to 12.6]) was significant (Table 2). The cervical LCR tests were significantly less specific than the cervical PCR tests (difference = –0.6 percentage points [95% CI, –1.2 to –0.1]), whereas urine LCR and PCR specificities were similar (difference = –0.2 percentage points [95% CI, –1.7 to 1.3]) (Table 3).

Head-to-head comparison of cervical versus urine speci-

mens when tested by NAATs. A head-to-head comparison of the sensitivities (Table 2) and specificities (Table 3) of NAATs performed on cervical specimens with those of the same NAATs performed on urine specimens was constructed using the alternate technology NAAT specimen and patient standards. For these estimates of performance, we used an independent reference standard that employed NAATs but not culture; e.g., we compared cervical with urine LCR using cervical PCR and cervical plus urine PCR as reference standards. In this example, patients were classified as infected if cervical PCR results were positive in the specimen standard and if cervical PCR or urine PCR results were positive in the patient standard.

We found that the differences in sensitivities between cervical and urine LCR and PCR tests depended upon which reference standard we used (Table 2). The difference between estimated cervical and urine test sensitivity levels was largest when only a cervical reference test was used (specimen standard). This difference was statistically significant only for the LCR sensitivity estimate (difference = 12.2 percentage points [95% CI, 5.7 to 18.7]). Adding a urine specimen to the cervical reference standard (patient standard) minimized the estimated

TABLE 3. Estimates of LCR, PCR, and PACE 2 specificities for detection of *C. trachomatis* infection by type of specimen and reference standard

Standard, test, and specimen type(s)	No. of subjects uninfected	Specificity with ^a :				
		Cervical swab			Urine	
		LCR	PCR	PACE 2	LCR	PCR
One-specimen						
Culture, cervical swab	3,191	97.5 (96.4–98.7) ^b	98.2 (97.4–99.1) ^b	99.3 (99.0–99.6)	96.6 (94.6–98.6) ^b	97.0 (95.6–98.3) ^b
LCR, cervical swab	3,125		99.5 (99.2–99.8)	99.6 (99.4–99.9)		98.2 (97.2–99.1) ^b
PCR, cervical swab	3,169	98.1 (97.5–98.7)		99.3 (99.0–99.6)	97.2 (95.6–98.8) ^b	
Two-specimen patient (both specimens negative)						
Culture, cervical and urethral swabs	3,144	98.1 (96.8–99.4) ^b	98.7 (98.0–99.5) ^b	99.5 (99.2–99.7)	97.8 (96.3–99.3) ^b	98.0 (96.8–99.2) ^b
LCR, cervical swab and urine	3,058		99.7 (99.6–99.9)	99.7 (99.5–99.9)		99.4 (99.0–99.8)
PCR, cervical swab and urine	3,089	99.0 (98.6–99.4)		99.6 (99.4–99.8)	98.9 (97.6–100.0) ^b	

^a Values are percentages. Values in parentheses are 95% confidence intervals.

^b *P* value for heterogeneity among research centers, <0.05. The estimates and 95% confidence intervals were derived from a random effects model (see Materials and Methods).

difference in sensitivity between cervical and urine tests by reducing the estimated sensitivity of the cervical test. The resulting sensitivity estimates were less than 80% for both cervical and urine PCR tests. Use of these NAAT-based reference standards also allowed us to estimate the overall sensitivity of cervical culture to be 81.5% using the cervical LCR standard, 72.2% using the cervical plus urine LCR standard, 85.1% using the cervical PCR standard, and 74.7% using the cervical plus urine PCR standard.

The differences in specificity between cervical and urine LCR and PCR tests also depended upon which reference standard was used, but the effects on absolute specificity were in the opposite direction from those described above for the comparison of the sensitivities of these tests (Table 3). Cervical and urine LCR and PCR specificity estimates increased when we substituted a cervical swab and urine specimen standard (patient standard) for a cervical swab standard (specimen standard), and the increase was greatest for the urine specificity estimate. Using the patient standard, LCR cervical and urine specificity estimates were similar (difference = 0.05 percentage points [95% CI, -1.1 to 1.2]) and did not exceed 99.0%. In contrast, the estimate for PCR cervical specificity was significantly higher (99.7%) than that of PCR urine specificity (99.4%) using the patient standard (difference = 0.3 percentage points [95% CI, 0.02 to 0.7]).

Variation in test performance among research centers. Tests for heterogeneity among the five research centers were performed for all estimates of test performance (Tables 2 and 3). Significant variation (or heterogeneity) was particularly notable for sensitivity estimates for PACE 2, urine LCR, and urine PCR tests. With one exception, the results of tests of heterogeneity of cervical LCR and PCR sensitivity estimates were not significant, regardless of which reference standard was used; the widest 95% CI range covered 11.3 percentage points. The exception was for cervical LCR sensitivity estimates when the cervical plus urethral culture patient standard was used ($P = 0.038$). By contrast, heterogeneity among research centers for estimates of PACE 2 sensitivity was significant for most of the reference standards used; the range covered by the 95% CIs was 10.9 to 20.4 percentage points. Estimates of urine LCR and PCR sensitivities were at least as heterogeneous as the PACE 2 sensitivity estimates, regardless of the reference standard employed.

The results of tests for heterogeneity of the PACE 2 specificity estimates were not significant, regardless of the reference standard used. Estimates of cervical and urine LCR and PCR specificity were heterogeneous when we used either the culture specimen standard or the culture patient standard. However, tests for heterogeneity of cervical LCR and PCR specificity estimates were not significant when we used the alternate NAAT as the reference standard with or without a urine specimen as part of the standard. Tests for heterogeneity of urine LCR specificity estimates were significant whether we used cervical PCR or cervical PCR plus urine PCR as reference standards. Estimates of urine PCR specificity were significantly heterogeneous when we used cervical LCR as the standard (95% CI ranged over 1.9 percentage points) but not significant when we expanded the LCR standard to include urine (95% CI ranged over 0.8 percentage points).

Substitution of NAATs for culture and use of a patient

standard for an improved reference standard. The sensitivities (Table 2) and specificities (Table 3) of PCR and LCR estimated with culture and alternate NAAT specimen and patient standards were determined using cervical swabs and urine specimens. The effect of adding a urethral swab or urine specimen to each of the culture-based, LCR-based, and PCR-based reference standards is shown in Tables 2 and 3. In addition, the effect on estimates of performance of tests exclusively using NAAT-technology-based reference standards that do not include any culture tests is illustrated in Tables 2 and 3. Adding urethral to cervical culture specimens as the reference standard modestly decreased estimates of cervical, but not urine, test sensitivities (Table 2). Similarly, adding urine NAAT to cervical NAAT as the reference standard decreased estimates of cervical and urine test sensitivities, but this effect was much more pronounced for the cervical tests than for urine tests (Table 2). Substituting cervical LCR for a cervical culture standard also decreased estimates of test sensitivities. The decreases were smaller for urine than cervical tests. Substituting cervical PCR for a cervical culture standard was associated with smaller decreases in estimates of cervical test sensitivities and no decrease for urine test estimates. Estimates of cervical test sensitivity were substantially reduced when both a urine specimen was added to the reference standard and LCR or PCR was substituted for the culture reference test. This combined effect was less for urine than cervical tests. When a NAAT was substituted for culture as the reference standard, there was a greater decrease in the sensitivity estimate for PACE 2 than for cervical PCR or LCR. Consequently, the proportional increase in numbers of infections detected by PCR or LCR relative to PACE 2 was greater when a NAAT was used than when culture was used as the reference standard (25 versus 17% infections detected by LCR versus PACE 2, respectively; 38 versus 28% infections detected by PCR versus PACE 2, respectively).

PACE 2 specificity estimates were higher than those for LCR and PCR when the reference standard was based on culture (Table 3). The increase in PACE 2 specificity associated with adding urethral or urine specimens to cervical specimens, or substituting LCR or PCR for culture as the reference standard, was small. Specificity estimates were lowest when cervical culture was used as the reference standard (specimen standard) and generally increased when urethral swab-urine specimens were added to a cervical reference standard and when a NAAT was substituted for culture as the reference test (Table 3). The increases were greater for estimates of urine test specificity than for cervical test specificity and greater for LCR and PCR than for PACE 2. Substituting LCR for culture as the reference standard was associated with a larger increase in estimates of test specificity than substituting PCR for culture.

DISCUSSION

NAATs are generally considered to be substantially more sensitive than antigen detection or nonamplified nucleic acid probe tests. Most studies have compared NAATs with culture and have employed discrepant analysis to obtain sensitivity estimates for NAATs and culture. Compared in this way, NAAT sensitivity estimates have nearly always exceeded the

those of the sensitivity of culture. The presumption that NAATs are superior to other nonculture, nonamplified tests is based largely on the premise that culture, when performed well, is at least as sensitive as any nonamplification test. Discrepant analysis involves using additional reference tests to classify subjects with regard to infection status whenever the evaluated test is positive but the reference standard is negative, a circumstance that occurs most often when the evaluated test is more sensitive than the reference standard. However, concern has been raised about the potential for exaggerated estimates of sensitivity and specificity due to selection bias, because not all tests are applied to all specimens (13, 19, 20). Relatively few studies have compared the amplified and nonamplified tests directly, and we know of no published studies that have made direct comparisons while avoiding discrepant analysis as we have in this analysis.

Our study, using a head-to-head comparative design and an independent reference standard, demonstrates that commercial LCR and PCR are substantially more sensitive than a widely used, nonamplified DNA probe test, detecting 17 to 38% more infections. The PCR sensitivity estimates were consistently, but not significantly, lower than the LCR estimates in this study, regardless of the source of clinical specimen. We evaluated the first-generation uniplex PCR in this study, modified by replacing the manufacturer's swab transport medium with a universal transport medium. A modification of this assay in a multiplex format has replaced the uniplex assay we tested (9, 22).

All of the tests studied exhibited high specificities. The specificities of commercial PCR and LCR have approached 100% in numerous studies employing discrepant analysis (3, 13). In contrast, the specificities of the NAATs in this study that were based on a reference standard consisting only of culture were less than 100% but were underestimated due to false-negative culture results. An example of this is our misleading result that PACE 2 specificity was higher than PCR and LCR specificity when the standard was culture alone. Using the alternate-technology NAAT on both cervical and urine specimens as the reference standard, e.g., cervical plus urine LCR as a standard for evaluating PCR, resulted in higher estimates of specificity that were still lower than published estimates, primarily because of variability among centers. This was especially true for estimates of LCR specificity. There are several possible explanations for this observation. First, we did not use discrepant analysis in this study, though the effect on the comparison would have been minimal, as the bias towards falsely high estimates of specificity has been shown to be small (11). Second, the lower estimates for LCR specificity may have been due to false-negative PCR tests, as this test appears to be somewhat less sensitive than LCR. In this circumstance, the bias is toward falsely low specificity of the more sensitive test. Finally, the lower specificity estimates for LCR observed in this study may represent real-world conditions reflecting laboratory-to-laboratory variation in performance of one or more of the assays that is not apparent in studies involving only a few or a single study center.

C. trachomatis can infect multiple anatomic sites. Several recent studies have reported that cervical specimens result in a higher sensitivity than urine specimens when both are tested by a NAAT (5, 7, 8, 29). However, all of these studies employed

a cervical test as a reference standard. It was previously demonstrated that *C. trachomatis* cervical test sensitivity is overestimated with regard to the patient when the reference standard includes only cervical culture (24). As expected, our results demonstrate that the bias in sensitivity estimates is more severe for cervical than urine tests. We found that the cervical and urine LCR and PCR tests are all nearly equivalent in sensitivity when we used a patient standard, although there was a trend toward slightly higher sensitivities for cervical compared with urine tests. This result suggests that, although the difference may be small, endocervical swab specimens are superior in performance to urine specimens for diagnostic tests for *C. trachomatis*. This finding supports a recommendation that, whenever possible, a cervical swab specimen, rather than urine, should be collected for *C. trachomatis* testing in women. However, urine specimens perform well with the NAATs, have greatly expanded the potential for screening both men and women, and should be collected for testing in settings where genital examinations are not feasible. Since either specimen could potentially result in a false-negative result, collection of both endocervical and urine specimens for testing individually or in combination may improve overall performance. Vaginal swab specimens, although quite promising in recent studies (4, 14, 26, 27), were not included for comparison in this study.

The design of studies evaluating diagnostic tests for *C. trachomatis* and selection of appropriate reference standards has been difficult and lacks standardization due to the cost and complexity of applying the historical "gold standard," culture, and the dilemma of newer technologies being more sensitive than culture, especially when applied to urine specimens. LCR and PCR are candidates to replace culture as reference tests, since they are more sensitive and better standardized and have had equivalent specificity in most studies (7, 16, 22, 28, 29). An additional advantage of using NAATs is the ability to incorporate a highly sensitive urine test into a two-test patient standard (1), thereby reducing overestimation of cervical test sensitivity and underestimation of cervical and urine test specificities that can result from use of a single-test specimen standard. Reducing these biases is especially important when comparing cervical and urine tests. Despite use of this improved reference standard, we demonstrated a trend toward greater sensitivity of cervical compared with urine tests.

As a new model for design of test evaluation studies, we recommend including a urine test in the reference standard to more nearly approximate a patient standard and to reduce overestimation of cervical test sensitivity. When comparing two cervical tests, a head-to-head comparison using a cervical standard is the best way of comparing how well the two tests detect a cervical infection. On the other hand, use of a patient standard is important because it reveals how a specific test performs in identifying a patient who is infected only at the urethral or at both the urethral and cervical sites. Thus, for overall estimates of test performance and when comparing cervical and urine tests, both cervical and urine tests should be included in the reference standard. Reference standards should not include the test(s) under evaluation. Culture tests are not necessary as a part of the standard and can be replaced with any of the commercial NAATs. Exclusion of culture from the standard will avoid the problem of overestimation of cervical test sensitivity due to false-negative culture results and facilitate

comparison of different studies, since the commercial tests are better standardized than culture methods. An example of this type of standard would be to use cervical plus urine PCR or LCR tests to evaluate a new NAAT. Patients who are either cervical or urine positive by one of the FDA-cleared NAATs would be classified as infected, while patients who are both cervical and urine negative by the FDA-cleared NAATs would be classified as uninfected. Since *C. trachomatis* infects multiple anatomic sites, use of the independent patient standard will also improve evaluation of alternative specimens such as vaginal or perineal swabs or tampons.

Lastly, it is important to note that our results varied substantially among centers. Differences in culture performance probably contributed to the interlaboratory variation in performance estimates of the evaluated tests (23). As noted above, reducing variation in reference test performance as a source of error is an important reason to consider substitution of NAATs for culture as reference tests. However, the magnitude of interlaboratory variation in performance estimates that we obtained for NAATs when we used the alternate NAAT as the reference standard was unexpected, given that these commercial tests are better standardized than culture. Interlaboratory variation was greater for urine than cervical specimens. The source of the variation warrants further investigation but may be due to differences in urine processing procedures, e.g., inhibition by residual urine after pelleting or loss of target by unintended aspiration of part of the urine pellet (D. H. Martin and C. Cammarata, Abstr. 13th Meet. Internat. Soc. Sex. Transm. Dis. Res., abstr. 385, 1999). These findings strongly suggest that commercial NAATs may not always perform optimally, even in laboratories with highly experienced technicians. However, one of the benefits of a multicenter trial is the identification of unexpected variations in test performance. If identified, sources of variation may be correctable through modification of manufacturer's protocols. Quality assurance programs that include proficiency testing protocols are thus critical to successful use of these tests for accurate results, particularly in low-prevalence populations, for whom there is a greater risk for false-positive results.

REFERENCES

- Alonzo, T. A., and M. S. Pepe. 1999. Using a combination of reference tests to assess the accuracy of a new diagnostic test. *Stat. Med.* **18**:2987-3003.
- Barnes, R. C. 1989. Laboratory diagnosis of human chlamydial infections. *Clin. Microbiol. Rev.* **2**:119-136.
- Black, C. M. 1997. Current methods of laboratory diagnosis of *Chlamydia trachomatis* infections. *Clin. Microbiol. Rev.* **10**:160-184.
- Carder, C., A. J. Robinson, C. Broughton, J. M. Stephenson, and G. L. Ridgeway. 1999. Evaluation of self-taken samples for the presence of genital *Chlamydia trachomatis* infection in women using the ligase chain reaction assay. *Int. J. STD AIDS* **10**:776-779.
- Carroll, K. C., W. E. Aldeen, M. Morrison, R. Anderson, D. Lee, and S. Mottice. 1998. Evaluation of the Abbott LCx ligase chain reaction assay for detection of *Chlamydia trachomatis* and *Neisseria gonorrhoeae* in urine and genital swab specimens from a sexually transmitted disease clinic population. *J. Clin. Microbiol.* **36**:1630-1633.
- Centers for Disease Control and Prevention. 2000. Sexually Transmitted Disease Surveillance, September 2001. U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, Atlanta, GA.
- Chernesky, M. A., S. Chong, D. Jang, K. Luinstra, J. Sellors, and J. B. Mahony. 1997. Ability of commercial ligase chain reaction and PCR assays to diagnose *Chlamydia trachomatis* infections in men by testing first-void urine. *J. Clin. Microbiol.* **35**:982-984.
- Crotchfelt, K. A., L. E. Welsh, D. DeBonville, M. Rosenstrauss, and T. C. Quinn. 1997. Detection of *Neisseria gonorrhoeae* and *Chlamydia trachomatis* in genitourinary specimens from men and women by a coamplification PCR assay. *J. Clin. Microbiol.* **35**:1536-1540.
- DiDomenico, N., H. Link, R. Knobel, T. Caratsch, W. Weschler, Z. G. Loewy, and M. Rosenstrauss. 1996. COBAS AMPLICOR: fully automated RNA and DNA amplification and detection system for routine diagnostic PCR. *Clin. Chem.* **42**:1915-1923.
- Fleiss, J. L. 1981. Statistical methods for rates and proportions, 2nd ed. John Wiley & Sons, Inc., New York, N.Y.
- Green, T. A., C. M. Black, and R. E. Johnson. 1998. Evaluation of bias in diagnostic-test sensitivity and specificity estimates computed by discrepant analysis. *J. Clin. Microbiol.* **36**:375-381.
- Hadgu, A. 1996. The discrepancy in discrepant analysis. *Lancet* **348**:592-593.
- Hadgu, A. 1997. Bias in the evaluation of DNA-amplification tests for detecting *Chlamydia trachomatis*. *Stat. Med.* **16**:1391-1399.
- Hook, E. W., K. Smith, C. Mullen, J. Stephens, L. Rinehardt, M. S. Pate, and H. H. Lee. 1997. Diagnosis of genitourinary *Chlamydia trachomatis* infections by using the ligase chain reaction on patient-obtained vaginal swabs. *J. Clin. Microbiol.* **35**:2133-2135.
- Jang, D., J. W. Sellors, J. B. Mahony, L. Pickard, and M. A. Chernesky. 1992. Effects of broadening the gold standard on the performance of a chemiluminometric immunoassay to detect *Chlamydia trachomatis* antigens in centrifuged first void urine and urethral swab samples from men. *Sex. Transm. Dis.* **19**:315-319.
- Johnson, R. E., T. A. Green, J. Schachter, et al. 2000. Evaluation of nucleic acid amplification tests as reference tests for *Chlamydia trachomatis* infections in asymptomatic men. *J. Clin. Microbiol.* **38**:4382-4386.
- Laird, N. M., and F. Mosteller. 1990. Some statistical methods for combining experimental results. *Int. J. Technol. Assessment Health Care* **6**:5-30.
- Lefebvre, J., H. Laperriere, H. Rousseau, and R. Masse. 1988. Comparison of three techniques for detection of *Chlamydia trachomatis* in endocervical specimens from asymptomatic women. *J. Clin. Microbiol.* **26**:726-731.
- McAdam, A. J. 2000. Discrepant analysis: how can we test a test? *J. Clin. Microbiol.* **38**:2027-2029.
- Miller, W. C. 1998. Bias in discrepant analysis: when two wrongs don't make a right. *J. Clin. Epidemiol.* **51**:219-231.
- Moore, D. S., and G. P. McCabe. 1989. Introduction to the practice of statistics. W. H. Freeman and Company, New York, N.Y.
- Pasternack, R., P. Vuorinen, T. Pitkajarvi, M. Koskela, and A. Miettinen. 1997. Comparison of manual Amplicor PCR, Cobas Amplicor PCR, and LCx assays for detection of *Chlamydia trachomatis* infection in women by using urine specimens. *J. Clin. Microbiol.* **35**:402-405.
- Pate, M. S., and E. W. Hook. 1995. Laboratory to laboratory variation in *Chlamydia trachomatis* culture practices. *Sex. Transm. Dis.* **22**:322-326.
- Schachter, J., J. Moncada, R. Whidden, et al. 1995. Noninvasive tests for diagnosis of *Chlamydia trachomatis* infections: application of ligase chain reaction to first-catch urine specimens of women. *J. Infect. Dis.* **172**:1411-1414.
- Schachter, J., W. E. Stamm, and T. C. Quinn. 1998. Discrepant analysis and screening for *Chlamydia trachomatis*. *Lancet* **351**:217-218.
- Stary, A., B. Najim, and H. H. Lee. 1997. Vulval swabs as alternative specimens for ligase chain reaction detection of genital chlamydial infection in women. *J. Clin. Microbiol.* **35**:836-838.
- Thomas, B. J., T. Pierpoint, D. Taylor-Robinson, and A. M. Renton. 1998. Sensitivity of the ligase chain reaction assay for detecting *Chlamydia trachomatis* in vaginal swabs from women who are infected at other sites. *Sex. Transm. Infect.* **74**:140-141.
- Van der Pol, B., D. V. Ferrero, L. Buck-Barrington, E. W. Hook III, C. Lenderman, T. Quinn, C. A. Gaydos, J. Lovchik, J. Schachter, J. Moncada, G. Hall, M. J. Tuohy, and R. B. Jones. 2001. 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.* **39**:1008-1016.
- van Doornum, G. J. J., L. M. Schouls, A. Pijl, I. Cairo, M. Buimer, and S. Bruisten. 2001. Comparison between the LCx probe system and the COBAS AMPLICOR system for detection of *Chlamydia trachomatis* and *Neisseria gonorrhoeae* infections in patients attending a clinic for treatment of sexually transmitted diseases in Amsterdam, The Netherlands. *J. Clin. Microbiol.* **39**:829-835.