CLINICAL AND VACCINE IMMUNOLOGY, Jan. 2008, p. 106–114 Vol. 15, No. 1
1556-6811/08/$08.00+0 doi:10.1128/CVI.00223-07

Utility of Composite Reference Standards and Latent Class Analysis in Evaluating the Clinical Accuracy of Diagnostic Tests for Pertussis▽

Andrew L. Baughman,[1]* Kristine M. Bisgard,[2] Margaret M. Cortese,[1] William W. Thompson,[3] Gary N. Sanden,[1]† and Peter M. Strebel[1]

National Center for Immunization and Respiratory Diseases,[1] Office of Workforce and Career Development,[2] and Office of the Director,[3] Centers for Disease Control and Prevention, Atlanta, Georgia

Received 31 May 2007/Returned for modification 28 August 2007/Accepted 25 October 2007

Numerous evaluations of the clinical sensitivity and specificity of PCR and serologic assays for *Bordetella pertussis* have been hampered by the low sensitivity of culture, the gold standard test, which leads to biased accuracy estimates. The bias can be reduced by using statistical approaches such as the composite reference standard (CRS) (e.g., positive if culture or serology positive; negative otherwise) or latent class analysis (LCA), an internal reference standard based on a statistical model. We illustrated the benefits of the CRS and LCA approaches by reanalyzing data from a 1995 to 1996 study of cough illness among 212 patients. The accuracy of PCR in this study was evaluated using three reference standards: culture, CRS, and LCA. Using specimens obtained 0 to 34 days after cough onset, estimates of the sensitivity of PCR obtained using CRS (47%) and LCA (34%) were lower than the culture-based estimate (62%). The CRS and LCA approaches, which utilized more than one diagnostic marker of pertussis, likely produced more accurate reference standards than culture alone. In general, the CRS approach is simple, with a well-defined disease status. LCA requires statistical modeling but incorporates more indicators of disease than CRS. When three or more indicators of pertussis are available, these approaches should be used in evaluations of pertussis diagnostic tests.

Despite the availability of an effective pertussis vaccine since the mid-1940s, pertussis (whooping cough) remains endemic in the United States. In 2004, a total of 25,827 pertussis cases were reported to the Centers for Disease Control and Prevention (CDC) (24). Adolescents and adults accounted for the majority (67%) of reported cases. Laboratory diagnosis of pertussis is particularly difficult in these age groups, thereby limiting detection and control.

For patients suspected of having pertussis, two types of clinical samples can be tested: a nasopharyngeal (NP) specimen for the isolation of *Bordetella pertussis* or for a PCR assay for *B. pertussis* DNA and a serum sample for the measurement of antibodies to *B. pertussis* antigens (11).

B. pertussis isolation by microbial culture is the conventional gold standard for confirming pertussis (37, 60). Most studies have derived sensitivity and specificity estimates of PCR or serologic tests using culture results as the gold standard (37). Sensitivity is the proportion of the true diseased patients classified as positive, and specificity is the proportion of the true nondiseased patients classified as negative (62). However, culture is an insensitive test, because the organism is fastidious and often not recoverable from the nasopharynx more than 3 weeks after cough onset (37). Because culture has low sensitivity, it cannot be used to determine the true specificity of other or new pertussis tests. Both PCR assays for *B. pertussis* DNA and serologic assays for antibodies to *B. pertussis* antigens have not been standardized, and their sensitivity and specificity are incompletely defined (27, 37, 41, 60).

Consider a diagnostic test under investigation, hereafter referred to as the index test. If culture for pertussis is assumed to be <100% sensitive and 100% specific and culture is used as the gold standard for assessing the index test, then the index test's sensitivity estimate will be unbiased but the specificity estimate will be biased in the direction of lower estimates (38, 46, 53, 62). This bias, referred to as the imperfect gold standard bias (62), occurs because some index test-positive results from truly infected persons will have been falsely negative by culture. Under the assumption that the index test and culture are conditionally independent, the negative bias of the specificity estimate increases as the sensitivity of culture decreases and as the prevalence of pertussis increases (46) (Fig. 1).

Biased estimates of an index test's performance parameters can have a substantial impact on patient management, public health response, and epidemiologic research (21). In general, the magnitude and direction of the potential bias in sensitivity and specificity based on an imperfect gold standard depend on whether the index test and gold standard tend to misclassify the same patients. When the classification errors caused by the index test and gold standard are independent, the estimates of sensitivity and specificity of the index test will be less than their true values (38, 62) (Fig. 2). However, when the classification errors caused by the index test and gold standard are dependent, the estimates of sensitivity and specificity can be biased in either direction (38, 50, 62). These patterns of bias can be illustrated by pointwise nonsampling intervals for sensitivity and specificity that reflect the maximum possible values of these parameters due to misclassification (Fig. 2). If the classification errors tend to occur in the same patients (positive dependence), then sensitivity and specificity will be overestimated (51, 53).

* Corresponding author. Mailing address: Centers for Disease Control and Prevention, 1600 Clifton Road, NE, Mailstop C-25, Atlanta, GA 30329. Phone: (404) 639-8198. Fax: (404) 639-2483. E-mail: ALB1@cdc.gov.

† Present address: 1830 Mountain Valley Rd., La Veta, CO 81055.

▽ Published ahead of print on 7 November 2007.

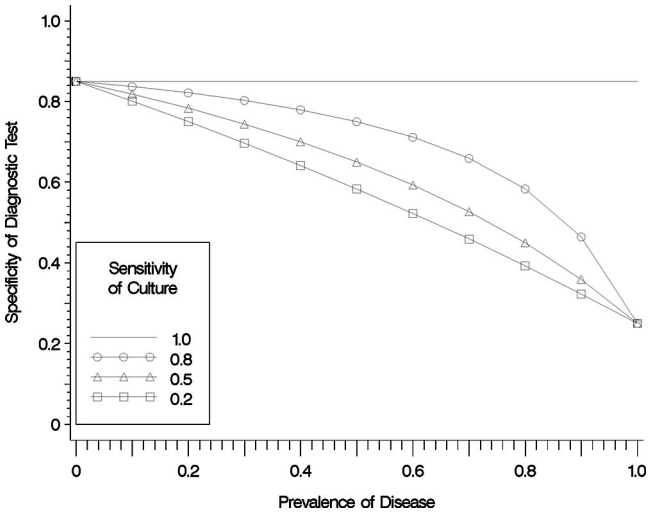

FIG. 1. Example of the observed specificity of a diagnostic test versus the prevalence of disease, by the sensitivity of culture, assuming a culture specificity of 100% and conditional independence between the diagnostic test and culture. Adjustment formulas for the observed specificity also assume a diagnostic test specificity of 85% and a diagnostic test sensitivity of 75% (46).

In the absence of a gold standard that has both high sensitivity and high specificity, few options exist to evaluate the accuracy of diagnostic tests. One option is to use a combination of several imperfect tests to define a better reference standard, a composite reference standard (CRS) (1). Another option is latent class analysis (LCA), which involves fitting a statistical model using all available diagnostic tests to define an internal reference standard (17, 23, 59). For both analytic approaches it is assumed that additional cases could be detected and considered true cases using the new reference standard.

Standardized PCR and serologic assays for diagnosing pertussis currently are under development (4, 10, 60). The evaluation of the accuracy of these assays will require a reference standard with both high sensitivity and specificity. In this report, we outline methods for the two alternative reference standards, CRS and LCA, illustrate them for evaluating a PCR assay by using data from a previously published study of pertussis in adolescents and adults, and make recommendations for their use.

MATERIALS AND METHODS

Several methods exist for reducing imperfect gold standard bias.

Discrepant analysis. In several evaluations of diagnostic tests for detecting *B. pertussis* infection, investigators have attempted to improve the sensitivity of culture by performing discrepant analysis (30, 57). A typical discrepant analysis involves the selective testing of the index test-positive, gold standard-negative specimens with a third, resolver test, which is considered a perfect gold standard (100% specific and 100% sensitive) (1, 52). If the resolver test result is positive, then it is considered a true positive (19); this reclassification leads to a modified or expanded gold standard that is based in part on the results of the index test (1). The incorporation of the results of the index test also can occur in other forms of discrepant analysis, in which an expanded gold standard is created by resolving discrepant specimens with patient histories (30). The incorporation of the results of the index test into the gold standard results, a phenomenon termed incorporation bias, typically overestimates the accuracy of the index test (39, 62). Thus, the selective testing or resolving in discrepant analysis merely substitutes incorporation bias for imperfect gold standard bias (35). To prevent incorporation bias, investigators advocate study designs in which all diagnostic tests are applied to each subject (18, 29, 31, 35, 36).

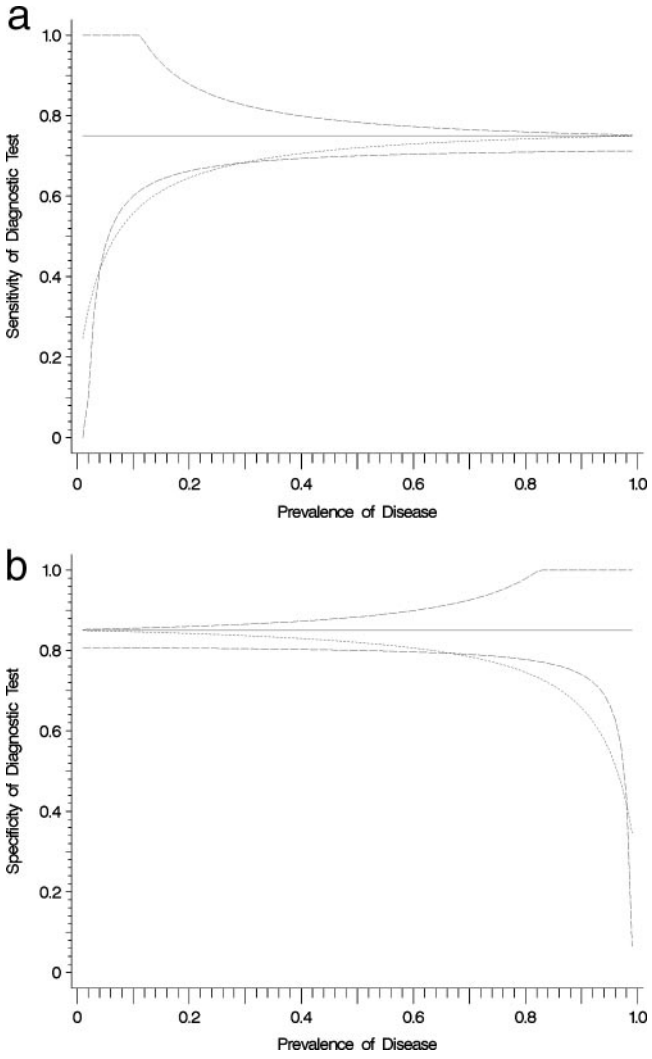

FIG. 2. (a) Example of the observed sensitivity of a diagnostic test versus the prevalence of disease, when the assumption of conditional independence between the diagnostic test and an imperfect gold standard may or may not be met. The solid line indicates the assumed level of sensitivity of the diagnostic test. The short-dashed line indicates the observed sensitivity assuming conditional independence calculated by adjustment formulas (46). The long-dashed lines indicate the pointwise nonsampling intervals for the observed sensitivity assuming diagnostic test-gold standard dependence. Formulas for the nonsampling intervals assume a diagnostic test sensitivity of 75%, a gold standard sensitivity of 95%, and a gold standard specificity of 95% (50). (b) Example of the observed specificity of a diagnostic test versus the prevalence of disease, when the assumption of conditional independence between the diagnostic test and an imperfect gold standard may or may not be met. The solid line indicates the assumed level of specificity of the diagnostic test. The short-dashed line indicates the observed specificity assuming conditional independence calculated by adjustment formulas (46). The long-dashed lines indicate the pointwise nonsampling intervals for the observed specificity assuming diagnostic test-gold standard dependence. Formulas for the nonsampling intervals assume a diagnostic test specificity of 85%, a gold standard sensitivity of 95%, and a gold standard specificity of 95% (50).

CRSs. To reduce imperfect gold standard bias but avoid incorporation bias, Hadgu (20) and Miller (35, 36) suggested combining multiple tests to improve

TABLE 1. Notation of variables used in a composite reference standard that combines culture with a resolver test to evaluate performance of a diagnostic test for pertussis[a]

Stage	Diagnostic test result	Variable when comparison test result was[b]: Positive		Negative	
		Notation	Data example	Notation	Data example
1	Positive	n_{++}	5	n_{+-}	5
	Negative	n_{-+}	3	n_{--}	199
	Total	$n_{++} + n_{-+}$	8	$n_{+-} + n_{--}$	204
2	Positive	n_{+-+}	2	n_{+--}	3
	Negative	n_{--+}	5	n_{---}	194
	Total	$n_{+-+} + n_{--+}$	7	$n_{+--} + n_{---}$	197

[a] See Table 2 for specificity and sensitivity formulas that use these variables.
[b] The comparison test of stage 1 is culture. Only culture-negative results are subject to stage 2 testing (the comparison test at stage 2 should be defined by investigators). Culture-negative results in stage 1 are classified according to the results of the resolver test in stage 2 (1). The data examples (values are numbers of samples) are from a previous study in which the CRS combined culture with IgG-PT to evaluate the accuracy of PCR in a previous study of 212 subjects (48).

the single best reference test. For example, a test with high specificity but poor sensitivity (e.g., culture) could be combined with another test with higher sensitivity (e.g., serology) to provide a relatively accurate CRS. The CRS can be formulated in the framework of a two-stage study design (1). In the first stage, all specimens are tested by the index test (e.g., PCR) and culture, but in the second stage, only those specimens that are culture negative are tested by the resolver test (e.g., serology) (Tables 1 and 2). The CRS is defined as positive if the specimen tested positive by either culture or the resolver test and is defined as negative if the specimen tested negative by both tests. The CRS also can be implemented in single-stage studies (33) and by combining more than two laboratory tests and clinical findings (28). Large-sample confidence intervals can be calculated for estimates of sensitivity and specificity (14).

The resolver test for forming a CRS can be chosen on the basis of the results of prior studies, clinical judgment, and an item analysis of all available diagnostic tests. Item analysis was developed to identify a final set of items (survey questions) in constructing a new psychological or achievement test (12). Item analysis can be adapted for diagnostic test evaluation to help determine which diagnostic test best discriminates between diseased and nondiseased persons. Consider a study design that involves J (≥ 3) diagnostic tests ($Y_1, Y_2, \ldots, Y_J$) rating each subject on a binary scale (1 = positive; 0 = negative). In an item analysis of diagnostic tests, the total score on all the diagnostic tests ($\sum Y_j$) is used as the operational definition of the disease likelihood level, which is considered to be a continuous variable. The association between the total score and the score on the jth diagnostic test (Y_j) is evaluated by the point biserial correlation (5). Higher values of the correlation indicate a stronger association between the disease likelihood level and a positive diagnostic test result.

LCA for assessing relative accuracy of diagnostic tests. A second valid approach for reducing imperfect gold standard bias is LCA. LCA is a mathematical correction that involves fitting a latent class model using data from all available diagnostic tests (32, 59). All diagnostic tests, including the gold standard, are regarded as imperfect. The latent class model assumes that for a randomly selected subject, the unobserved true state of disease, the latent variable (X), influences the observed measurements made by J diagnostic tests (Fig. 3a). The model makes two key assumptions, which are described below.

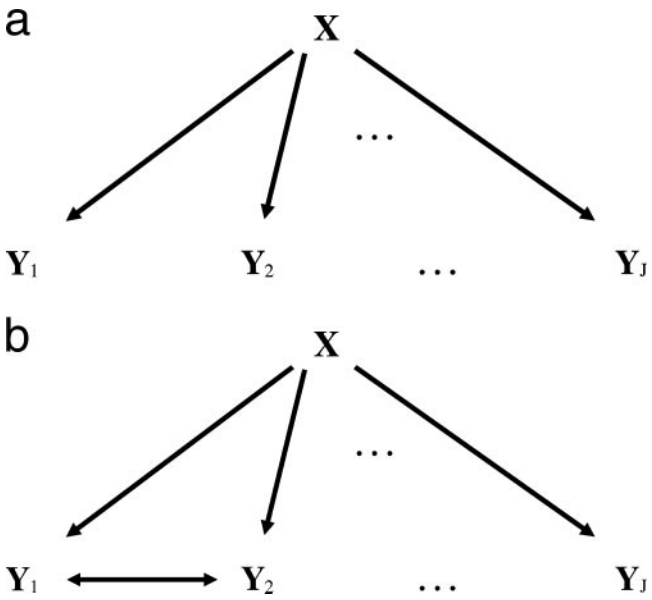

FIG. 3. (a) Conditional independence latent class model. The observed measurements made by diagnostic tests (Y_j; $j = 1, \ldots, J$) are independent, given the common latent variable for subject disease status (X). If an association is observed among diagnostic tests, then it is entirely attributable to the common factor X. (b) Example of a dependence latent class model that includes a bivariate association between Y_1 and Y_2. The observed association between Y_1 and Y_2 cannot be entirely explained by the common factor X.

Assumption 1. The study population consists of two internally homogeneous subpopulations or latent classes. These two latent classes are mutually exclusive and exhaustive and represent the latent diseased ($X = 1$) and nondiseased ($X = 0$) populations. Subjects in the same latent class are assumed to be homogeneous with respect to the likelihood of disease. For example, consider the observed two-by-two table defined by Y_1 and Y_2. The joint probability for the row-1, column-1 cell of the table can be factored by the rule of total probability (14):

$$P(Y_1 = 1, Y_2 = 1) = P(Y_1 = 1, Y_2 = 1|X = 1)P(X = 1) + P(Y_1 = 1, Y_2 = 1|X = 0)P(X = 0) \quad (1)$$

where $P(X = 1) + P(X = 0) = 1$.

Assumption 2. Y_j is independent of $Y_{j'}$, $j \neq j'$, given X. This is an assumption of conditional independence of the diagnostic tests, given the true disease status of the subject. This assumption means that an observed measurement made by a diagnostic test depends only on the true disease status of the subject and not

TABLE 2. Notation of formulas for calculating sensitivity and specificity of a composite reference standard that combines culture with a resolver test to evaluate performance of a diagnostic test for pertussis[a]

Stage	Formula for: Sensitivity		Specificity	
	Notation	Data example	Notation	Data example
1	$n_{++}/(n_{++} + n_{-+})$	$5/(5 + 3) = 0.625$	$n_{--}/(n_{--} + n_{+-})$	$199/(199 + 5) = 0.975$
2	$(n_{++} + n_{+-+})/(n_{++} + n_{+-+} + n_{-+} + n_{--+})$	$(5 + 2)/(5 + 2 + 3 + 5) = 0.467$	$n_{---}/(n_{---} + n_{+--})$	$194/(194 + 3) = 0.985$

[a] Table 1 lists individual variables used in sensitivity and specificity formulas. The data examples are from a previous study in which the CRS combined culture with IgG-PT to evaluate the accuracy of PCR in a previous study of 212 subjects (48).

TABLE 3. Item analysis for determining which indicator of pertussis to combine with culture for evaluation of PCR in a previous pertussis study of 212 subjects (48)

Indicator of pertussis and candidate for combining with culture	Positive		Point biserial correlation[a]
	No. of subjects	% of subjects	
Indicator			
Culture	8	3.8	NA
PCR	10	4.7	NA
Candidates			
IgG-PT	13	6.1	0.88
IgA-PT	6	2.8	0.65
IgG-FHA	21	9.9	0.83
IgA-FHA	16	7.5	0.80
Clinical case[b]	180	84.9	0.41

[a] Pearson product moment correlation between the indicator score (1 = positive; 0 = negative) and the total score for the seven indicators listed in this table (range, 0 to 7). NA, not applicable.
[b] Cough illness of ≥14 days' duration with one or more episodes of paroxysms of coughing, whoop, or posttussive vomiting.

on the measurements of other diagnostic tests or on any covariate (Fig. 3a). Thus, the diagnostic tests make classification errors independently of each other, irrespective of disease status. For the observed two-by-two table of Y_1 and Y_2, this assumption allows the joint probability in equation 1 to be factored further into an expression that includes conditional probabilities (a sensitivity parameter and a specificity parameter for each diagnostic test) and latent class probabilities (represented by a single prevalence parameter):

$$P(Y_1 = 1, Y_2 = 1) = P(Y_1 = 1|X = 1)P(Y_2 = 1|X = 1)P(X = 1) + P(Y_1 = 1|X = 0)P(Y_2 = 1|X = 0)P(X = 0) = Se_{Y_1}Se_{Y_2}\,\text{Prevalence} + (1 - Sp_{Y_1})(1 - Sp_{Y_2})(1 - \text{Prevalence}) \quad (2)$$

Under the two assumptions of this model, termed the conditional independence model, the joint probability or likelihood for J diagnostic tests can be similarly developed, and parameters can be estimated by maximum likelihood. The latent class model requires at least three diagnostic tests to have enough degrees of freedom in the observed data table (e.g., a two-by-two-by-two table) to estimate all of the parameters. Four or more diagnostic tests are required for models that include bivariate associations between diagnostic tests to account for conditional dependence (Fig. 3b). For example, culture and PCR may be positively dependent due to correlated classification errors among the same truly infected subjects; some infected subjects may be falsely negative for *B. pertussis* by both culture and PCR if too few bacteria exist in the nasopharynx (such as when an NP specimen is obtained >3 weeks after cough onset) or if the NP specimens are inadequately obtained or poorly handled, among other reasons (34).

The requirements for the number of diagnostic tests can be relaxed by restricting values of parameters to certain values or by restricting two or more parameters to have equal values (32, 59). The conditional independence model and models that account for conditional dependence between pairs of diagnostic tests can be fit using Latent GOLD software (55). A final model can be selected based on the value of the Bayesian information criterion, with smaller values representing better fits (32).

The results of fitting a latent class model to data from multiple diagnostic tests can be used to predict the disease status of individual study subjects (2, 3). Bayes' rule can be used to calculate the posterior probability that a subject is in the latent diseased population, given the subject's observed results on the multiple diagnostic tests.

Data example: prospective study of pertussis disease burden in adolescents and adults. A prospective study conducted among members of a managed care organization in Minneapolis/St. Paul, MN, measured the pertussis disease burden in adolescents and adults (48). Between January 1995 and December 1996, 212 persons aged 10 to 49 years who presented with an acute paroxysmal cough or a persistent cough illness of 7 to 34 days' duration were enrolled in the study. At enrollment, NP swab specimens were obtained for culture and PCR, and a first serum specimen was obtained. A second serum specimen was obtained ≥3 weeks later. Serum samples were assayed by an enzyme-linked immunosorbent assay for immunoglobulin G (IgG) and IgA against pertussis toxin (PT) and filamentous hemagglutinin (FHA), resulting in four types of serological test results: IgG-PT, IgA-PT, IgG-FHA, and IgA-FHA. The results of several laboratory tests were available: culture, PCR, and ≥2-fold increases in IgG-PT, IgA-PT, IgG-FHA, or IgA-FHA. Culture and PCR were performed in separate locations within the Minnesota Department of Health Laboratory, and the serologic assays were performed at Vanderbilt University (TN) (48). The method used for the conventional PCR assay was described by van der Zee et al. (54); primer pairs were based on insertion sequence elements IS*481* (specific for *B. pertussis*) and IS*1001* (specific for *B. parapertussis*). In addition to laboratory test results, classification by the pertussis clinical case definition was available (9). A clinical case was defined as a cough illness of ≥14 days' duration with one or more of the following symptoms: paroxysms of coughing (coughing spells with the inability to breathe during the spells), inspiratory whoop, or posttussive vomiting.

We analyzed these data by considering PCR to be the index test for illustrative purposes only. The sensitivity and specificity of PCR were estimated using culture results as the gold standard, a CRS, and LCA.

RESULTS

Of 212 study subjects, 180 (85%) subjects met the clinical case definition, 8 (3.8%) were culture positive, and 10 (4.7%) were PCR positive (Table 3). Of 13 subjects with positive results by culture or PCR, 5 subjects had positive results by both tests. The serology tests tended to identify a higher proportion of positive results (Table 3).

Culture results as the gold standard. When culture was used as the gold standard test, PCR had a sensitivity of 62% and a specificity of 98% (Table 4). IgG-PT had good sensitivity

TABLE 4. Sensitivity and specificity of indicators of pertussis in a previous pertussis study of 212 subjects (48) based on culture results, a CRS, and LCA

Indicator of pertussis	Culture[a]		CRS[b]		LCA[c]	
	Sensitivity (SE)	Specificity (SE)	Sensitivity (SE)	Specificity (SE)	Sensitivity (SE)	Specificity (SE)
Culture	NA	NA	NA	NA	0.339 (0.113)	0.990 (0.007)
IgG-PT	0.750 (0.153)	0.966 (0.013)	NA	NA	0.732 (0.110)	0.999 (0.001)
IgA-PT	0.250 (0.153)	0.980 (0.010)	0.400 (0.126)	1.000	0.338 (0.113)	0.999 (0.001)
IgG-FHA	0.750 (0.153)	0.926 (0.018)	0.867 (0.088)	0.959 (0.014)	0.971 (0.042)	0.980 (0.010)
IgA-FHA	0.625 (0.171)	0.946 (0.016)	0.733 (0.114)	0.975 (0.011)	0.843 (0.091)	0.995 (0.006)
Clinical case[d]	1.000	0.157 (0.026)	1.000	0.162 (0.026)	0.942 (0.057)	0.159 (0.026)
PCR	0.625 (0.171)	0.975 (0.011)	0.467 (0.129)	0.985 (0.009)	0.339 (0.113)	0.979 (0.010)

[a] Culture was the gold standard. NA, not applicable.
[b] The CRS was defined as positive if culture or IgG-PT was positive; otherwise, the CRS was defined as negative.
[c] The latent class model included all seven indicators of pertussis and a bivariate association between culture and PCR.
[d] Cough illness of ≥14 days' duration with one or more episodes of paroxysms of coughing, whoop, or posttussive vomiting.

TABLE 5. LCA of seven indicators of pertussis in a previous pertussis study of 212 subjects (48)[a]

Response pattern[b]	Observed frequency	Expected frequency	Standardized residual[e]	Posterior probability of pertussis[c]	Predicted latent class[d]
1111111	1	0.85	0.16	1.000	1
1110111	2	1.67	0.26	1.000	1
1110101	1	0.31	1.24	1.000	1
1100001	1	0.43	0.87	0.008	0
1011111	1	0.29	1.30	1.000	1
1010111	1	0.58	0.56	1.000	1
1000001	1	1.21	−0.20	0.001	0
0111111	1	0.29	1.30	1.000	1
0110101	1	0.11	2.72	0.999	1
0100001	3	2.85	0.09	0.000	0
0011111	3	1.94	0.76	1.000	1
0010111	2	3.80	−0.92	1.000	1
0000111	3	1.41	1.34	0.988	1
0000110	1	0.09	3.04	0.964	1
0000101	4	3.34	0.36	0.078	0
0000011	1	0.89	0.12	0.047	0
0000001	154	154.7	−0.06	0.000	0
0000000	31	29.3	0.31	0.000	0

[a] The latent class model included all seven indicators of pertussis and a bivariate association between culture and PCR.

[b] Response pattern represents the cross-classification of the seven binary indicators (1 = positive; 0 = negative) in the following order: culture, PCR, IgG-PT, IgA-PT, IgG-FHA, IgA-FHA, and clinical case definition.

[c] Posterior probability of pertussis was derived by Bayes' rule using latent class model parameter estimates.

[d] Predicted latent class was defined as diseased (class 1) if the posterior probability of pertussis was >0.5; otherwise, the predicted latent class was nondiseased (class 0) (3).

[e] The standardized residual was determined with the formula $(O - E)/\sqrt{E}$, where O is the observed frequency and E is the expected frequency.

(75%) and specificity (97%); none of the other serological tests performed as well as IgG-PT on both sensitivity and specificity. The pertussis clinical case definition had perfect sensitivity (100%) but was not specific (16%).

CRS. Given the known low sensitivity of culture, an item analysis was performed to help determine which indicator of pertussis to combine with culture results in a CRS. For this analysis, the six laboratory tests and the clinical case definition for pertussis were used to create a total score (range, 0 to 7). IgG-PT had the highest correlation with the total score (0.88) (Table 3). The correlations with the total score for IgG-FHA (0.83) and IgA-FHA (0.80) were relatively high; the lowest correlation occurred with the clinical case definition (0.41).

On the basis of these results and considering that PT is the most specific *B. pertussis* antigen tested (11), IgG-PT was chosen as the resolver test and was combined with culture in a CRS. Based on the CRS, the sensitivity of PCR was 47%, which was lower than that based on culture (62%), but the specificity of the two tests was the same (98%) (Table 4). The estimates of sensitivity and specificity based on the CRS for the other serology tests were higher than those based on culture.

LCA. An LCA was performed based on the cross-classification of results for all six laboratory tests and the clinical case definition (Table 5). Of 27 (12.7%) patients with at least one positive laboratory test result, all except one patient met the clinical case definition; this patient was positive only for IgG-FHA and IgA-FHA. After the conditional independence model was fitted, a large (statistically significant) bivariate residual was found for the culture-PCR association, so a second model that added a separate parameter for conditional dependence between culture and PCR was fitted. This model had a lower value of the Bayesian information criterion than the conditional independence model, and none of its bivariate residuals was large, suggesting a good fit of the model. The parameter estimates for the model were virtually identical to those for the conditional independence model. Based on the results of the conditional dependence model, PCR had a sensitivity of 34% and a specificity of 98% (Table 4). Culture and IgA-PT had relatively low sensitivity (34%), whereas the clinical case definition (94%) and IgG-FHA (97%) had relatively high sensitivity. All indicators except the case definition were relatively specific. IgG-PT and IgA-PT had the highest specificity (99.9%). The latent class model also provided an estimate of the prevalence of pertussis based on all seven indicators of pertussis (estimate, 8.4%; standard error, 1.9%).

For each observed response pattern, the results of the latent class model were used to calculate a posterior probability of pertussis (Table 5). If the posterior probability was >0.50, then subjects with that observed response pattern were classified as diseased (3). A total of 17 subjects were thus classified as diseased based on the latent class model, whereas 8 subjects were positive by culture and 15 subjects were positive by CRS. Two of the eight culture-positive subjects, one with a PCR-negative result and both with all serologic assays negative, had very low posterior probabilities of pertussis and were classified as nondiseased (Table 5). When the parameter for the specificity of culture was constrained to equal 1 in another latent class model (i.e., all eight culture-positive subjects were restricted to membership in the diseased latent class), all of the estimates of specificity for the other indicators of pertussis were virtually identical to the original estimates. The estimate of sensitivity changed only slightly for the PCR and clinical case definition, but it was higher for culture (a change from 34 to 38%) and was 6 to 12 percentage points lower for the four serology tests. The prevalence estimate increased from 8.4 to 10.0%.

Because the timing of specimen collection is a major determinant for culture, PCR, and antibody detection tests (41, 47, 48), we evaluated the performance of PCR within two subgroups of subjects defined by the interval between cough onset and enrollment (NP swab and first serum specimen): 0 to 13 days and 14 to 34 days (Table 6). With seven culture-positive samples in the 0- to 13-day interval but only one in the 14- to 34-day interval, the culture-based estimate of sensitivity was not reliable for the 14- to 34-day interval; the culture-based estimate of specificity did not vary by interval. However, results of both the CRS and LCA suggested that the sensitivity of PCR was significantly higher when the NP specimens were collected during the first 2 weeks of illness than when they were collected later in the course of illness (Table 6).

DISCUSSION

The laboratory diagnosis of pertussis is challenging, particularly for adolescents and adults, because the conventional gold standard, culture, is insensitive in these age groups (37). Significant research efforts recently have been devoted to developing a PCR and serologic assays to complement culture in

TABLE 6. Sensitivity and specificity of PCR in a previous pertussis study of 212 subjects (48) by the interval between cough onset and enrollment (NP swab and first serum specimen)

Interval (days)	Sample size	No. (%) of subjects PCR positive	Culture[a]		CRS[b]		LCA[c]	
			Sensitivity (SE)	Specificity (SE)	Sensitivity (SE)	Specificity (SE)	Sensitivity (SE)	Specificity (SE)
0–13	114	7 (6.1)	0.571 (0.187)	0.972 (0.016)	0.600 (0.155)	0.990 (0.010)	0.634 (0.156)	0.990 (0.010)
14–34	98	3 (3.1)	1.000	0.979 (0.014)	0.200 (0.179)	0.978 (0.015)	0.002 (0.014)	0.966 (0.019)
Total	212	10 (4.7)	0.625 (0.171)	0.975 (0.011)	0.467 (0.129)	0.985 (0.009)	0.339 (0.113)	0.979 (0.010)

[a] Culture was the gold standard.
[b] The CRS was defined as positive if culture or IgG-PT was positive; otherwise, it was defined as negative.
[c] The latent class model for each interval and for the total included all seven indicators of pertussis; however, the model for the 14- to 34-day interval and that for the total also included a bivariate association between culture and PCR.

diagnosing pertussis (41). Unless a gold standard with high sensitivity and specificity is used in evaluating a newly developed assay, the performance of the assay will be misrepresented and the consequent use of the assay will not be optimized.

CRSs and LCA are scientifically and statistically valid alternatives to using culture results as the gold standard in diagnostic accuracy studies. In our data example, they led to much lower estimates of sensitivity of PCR (47 and 34%, respectively) than the culture-based estimate (62%). The culture-based estimate of sensitivity was biased higher due to the positive dependence in classification by PCR and culture. We believe that the CRS and LCA provided a more accurate indicator of disease than culture results alone, because they defined disease status by using additional markers of disease. For instance, information on five subjects with positive IgG-PT results who tested negative by culture and PCR (Tables 1 and 2) was used to define pertussis in both the CRS and LCA approaches. In prior studies of pertussis in adolescents and adults that were based on case definitions that combined culture, PCR, and serologic results, it was found that the PCR sensitivity estimates were less than 50% (61).

The data example was limited by few positive results by culture ($n = 8$) or PCR ($n = 10$). For this reason, the results from our analysis on the performance of PCR as a diagnostic test may not be generally applicable, particularly in the period of 14 to 34 days after cough onset. In addition, the PCR and serologic assays used in the pertussis study were not standardized. These limitations are common to published evaluations of pertussis diagnostic tests. The CDC currently is working to optimize PCR for detecting *B. pertussis* DNA and also is working with the Food and Drug Administration to optimize serologic assays for *B. pertussis* antigens (10).

Several practical aspects of implementing the CRS approach deserve mention. An appropriate CRS for defining pertussis requires a highly sensitive resolver test to improve upon the low sensitivity of culture. The resolver test also must be specific, because the CRS assumes that the resolver test, as well as culture, is highly specific. Under these assumptions, the CRS will increase the sensitivity relative to that of culture and the resolver test but will remain highly specific (38). In the data example, the choice of IgG-PT as the resolver test was aided by performing an item analysis. The item analysis suggested that IgG-FHA and IgA-FHA also were good candidates for the resolver test. However, FHA antibody tests would not be good choices for the resolver test, because FHA antibodies also are observed in response to infection with other *Bordetella* species (4, 37). The two-stage framework of the CRS in which the resolver test was applied to all culture-negative results was used in the data example because of sparse data (Tables 1 and 2). Additional study designs for the CRS are based on sampling from the cells of the two-by-two table of the index test by culture results (1, 22).

The LCA approach is more complicated than the CRS approach, because it involves fitting a statistical model that models unobserved data, the latent disease status. Latent class models attempt to provide an approximation of the diagnostic truth based on the results of all available diagnostic tests, recognizing that the true classification of a person's disease status is unknown and can be defined only theoretically. Thus, a sensible analysis strategy would be to include in the model indicators of disease that are based on different biologic or physiologic phenomena. An additional advantage of latent class models over CRSs in this respect is that they do not require assumptions about the specificity of the resolver or any other test.

Latent class models have other advantages over the CRS approach. They allow the estimation of sensitivity and specificity of each diagnostic test included in the model. In addition, they provide estimates of disease prevalence, but it is important to note that the prevalence estimate depends on the patient sample as well as the particular variables included in the model. In the Minnesota pertussis study, 3.8% of the subjects were culture positive, and 12.7% had positive results on at least one laboratory test; the pertussis prevalence estimate was 8.4% based on the latent class model containing seven indicators of disease (Table 4). Lastly, latent class models may provide more reliable estimates of diagnostic performance than other approaches, because they model all of the available diagnostic information. In the subgroup analysis defined by the interval between cough onset and the first specimen collection, the latent class models indicated that, like culture, PCR was a more sensitive diagnostic test earlier in the patient's course of illness. This finding is consistent with results of previous pertussis studies (47, 48).

Although in the data example the latent class models were developed for the results of culture from all study subjects, latent class models also can be developed when culture is performed only for subjects with positive results on some other indicator(s) (58). However, they may require substantial

TABLE 7. Comparison of analysis approaches for evaluating the clinical accuracy of pertussis diagnostic tests

Analysis approach or reference standard	Sample of subjects assessed by each indicator of disease[a]	Comments	Recommendations
Culture (reference standard)	All subjects	Low sensitivity (12–60%) (60)	Define additional true positives by combining culture results with other indicators of disease using the CRS and LCA analysis approaches
Discrepant analysis	Selective testing of index test-positive/culture-negative subjects, or of all discordant results	Hard to justify or defend; leads to biased accuracy estimates (19, 20)	Not recommended
CRS	All subjects (1), or subjects sampled from cells in a two-by-two table of index test and culture (1, 22)	Simple approach; disease status is well defined	When an indicator of disease with high sensitivity and specificity is available, combine with culture results to form a CRS
LCA	All subjects (59), or reference test applied to index test-positive and other test(s)-positive subjects (58)	Statistical modeling required; incorporates more indicators of disease than CRS	Use when at least three indicators of disease are available; assess conditional independence assumption of model when four or more indicators are available
LCA with incorporation of prior information			
Assuming values of sensitivity/specificity for reference or other test(s)	All subjects (16, 59, 62)	Can use different sets of assumed values in a sensitivity analysis	Implement after a thorough review and summary of literature
Bayesian analysis, assuming a prior distribution for values of sensitivity/specificity of each test and prevalence of disease	All subjects (13, 15, 25)	More complicated and involves more analysis assumptions than non-Bayesian analysis	A good alternative for formally summarizing and incorporating results from prior studies

[a] Indicator of disease could be any diagnostic test or case definition. Index test is the diagnostic test under investigation.

amounts of data to avoid the estimation problems associated with the identifiability of models. A latent class model is identifiable if one solution to the likelihood equations exists. Even if the model is identifiable, the parameters may not be estimated uniquely because of a small sample size, sparseness in the observed data table, or an unusual pattern of observed data. To prevent boundary value problems and the nonexistence of maximum likelihood estimates, the Latent GOLD program uses Dirichlet priors with user-defined parameters for the latent and conditional response probabilities and a hybrid expectation-maximization/Newton-Raphson model-fitting algorithm (56). Recent applications of LCA have been useful for studies with sample sizes of about 150 subjects evaluated by four or five diagnostic tests (6, 8).

If one is willing to make further analysis assumptions about the clinical sensitivity and specificity values of the reference standard, then this information may be incorporated into an LCA (16, 59, 62). For example, when we assumed that culture was 100% specific and fit another latent class model for the 0- to 34-day interval, the accuracy estimates for PCR did not change, but the number of culture positives in the pertussis study was small. In other settings, this approach may be useful for determining the possible range of values for the index test under different assumptions about the accuracy of the reference standard. A more comprehensive approach to incorporating prior information on the accuracy of the reference standard would be to assume a prior distribution for sensitivity and specificity values in a Bayesian analysis (13, 15, 25).

Recommendations for evaluations of diagnostic tests. Table 7 compares all of the analysis approaches and provides recommendations for their use. Discrepant analysis should not be used in evaluating diagnostic tests for pertussis, because it violates a fundamental principle of diagnostic test evaluation: the reference standard should not incorporate any test result that depends on the results of the diagnostic test. The acceptance of this principle is challenging when evaluating a diagnostic test that truly is more sensitive than the conventional standard (culture) or even a perfect test; such a test will identify as diseased some individuals who were classified as nondiseased by the gold standard (42–45). Culture for *B. pertussis* has near-perfect specificity but poor sensitivity. PCR assays potentially can detect a single copy of *B. pertussis* target DNA (26), and prior pertussis studies have found that PCR assays, when performed with the appropriate laboratory control reagents, have high diagnostic sensitivity (37). However, they can yield false-positive results for several reasons, including clinical or laboratory contamination (7, 27, 34, 49). Numerous false-positive PCR results occurred in three recent outbreaks of respiratory illness mistakenly attributed to pertussis (10). False-negative results of PCR assays also are possible for a variety of reasons, not all of which can be rigorously controlled (34). Standardized serologic assays might be sensitive and specific enough to be utilized in diagnostic test evaluations, given that the variability inherent in these quantitative assays does not exceed the minimal levels for acceptance criteria (40).

No single laboratory test or clinical finding used to define

pertussis is perfect. One prudent analysis strategy for reducing imperfect gold standard bias is to combine serology, PCR, or clinical findings with culture results in a CRS; combining the results of additional markers of pertussis with culture results likely will provide a more accurate definition of disease. Because each laboratory test and clinical case definition adds different information about disease, LCA is attractive because it includes the results of all tests, including the index test.

In conclusion, the CRS and LCA approaches were useful for evaluating the relative accuracy of PCR and other diagnostic tests for pertussis in a prior pertussis study. These approaches likely provided more accurate reference standards than culture results alone and should be used in evaluations of diagnostic tests for pertussis. The best way forward is to ensure close interdisciplinary collaboration among clinicians, laboratorians, and statisticians.

ACKNOWLEDGMENTS

We are grateful to Freyja Lynn (National Institute of Allergy and Infectious Diseases, Bethesda, MD), Brian Plikaytis (CDC, Atlanta, GA), Trudy Murphy (CDC), and Lisa Cairns (CDC) for comments that improved the manuscript.

The findings and conclusions in this report are those of the authors and do not necessarily represent the views of the funding agency.

REFERENCES

1. **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.
2. **Alvord, W. G., J. E. Drummond, L. O. Arthur, R. J. Biggar, J. J. Goedert, P. H. Levine, E. L. Murphy, S. H. Weiss, and W. A. Blattner.** 1988. A method for predicting individual HIV infection status in the absence of clinical information. AIDS Res. Hum. Retrovir. **4:**295–304.
3. **Bartholomew, D. J., and M. Knott (ed.).** 1999. Latent class models, p. 133–156. *In* Latent variable models and factor analysis, 2nd ed. Oxford University Press, New York, NY.
4. **Baughman, A. L., K. M. Bisgard, K. M. Edwards, D. Guris, M. D. Decker, K. Holland, B. D. Meade, and F. Lynn.** 2004. Establishment of diagnostic cutoff points for levels of serum antibodies to pertussis toxin, filamentous hemagglutinin, and fimbriae in adolescents and adults in the United States. Clin. Diagn. Lab. Immunol. **11:**1045–1053.
5. **Bedrick, E. J.** 1998. Biserial correlation, p. 404–407. *In* P. Armitage and T. Colton (ed.), Encyclopedia of biostatistics, vol. 1. John Wiley & Sons, Inc., New York, NY.
6. **Boelaert, M., K. Aoun, J. Liinev, E. Goetghebeur, and P. Van der Stuyft.** 1999. The potential of latent class analysis in diagnostic test validation for canine *Leishmania infantum* infection. Epidemiol. Infect. **123:**499–506.
7. **Borst, A., A. T. A. Box, and A. C. Fluit.** 2004. False-positive results and contamination in nucleic acid amplification assays: suggestions for a prevent and destroy strategy. Eur. J. Clin. Microbiol. Infect. Dis. **23:**289–299.
8. **Butler, J. C., S. C. Bosshardt, M. Phelan, S. M. Moroney, M. L. Tondella, M. M. Farley, A. Schuchat, and B. S. Fields.** 2003. Classical and latent class analysis evaluation of sputum polymerase chain reaction and urine antigen testing for diagnosis of pneumococcal pneumonia in adults. J. Infect. Dis. **187:**1416–1423.
9. **Centers for Disease Control and Prevention.** 1990. Case definitions for public health surveillance. Morb. Mortal. Wkly. Rep. **39**(RR–13)**:**26–27.
10. **Centers for Disease Control and Prevention.** 2007. Outbreaks of respiratory illness mistakenly attributed to pertussis—New Hampshire, Massachusetts, and Tennessee, 2004–2006. Morb. Mortal. Wkly. Rep. **56:**837–842.
11. **Cherry, J. D., E. Grimprel, N. Guiso, U. Heininger, and J. Mertsola.** 2005. Defining pertussis epidemiology: clinical, microbiologic and serologic perspectives. Pediatr. Infect. Dis. J. **24:**S25–S34.
12. **Crocker, L., and J. Algina (ed.).** 1986. Item analysis, p. 311–338. *In* Introduction to classical and modern test theory. Harcourt Brace Jovanovich Inc., Orlando, FL.
13. **Dendukuri, N., and L. Joseph.** 2001. Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests. Biometrics **57:** 158–167.
14. **Fleiss, J. L., B. Levin, and M. C. Paik.** 2003. Statistical methods for rates and proportions, 3rd ed. John Wiley & Sons, Inc., New York, NY.
15. **Garrett, E. S., W. W. Eaton, and S. Zeeger.** 2002. Methods for evaluating the performance of diagnostic tests in the absence of a gold standard: a latent class model approach. Stat. Med. **21:**1289–1307.
16. **Gart, J. J., and A. A. Buck.** 1966. Comparison of a screening test and a reference test in epidemiologic studies. II. A probabilistic model for the comparison of diagnostic tests. Am. J. Epidemiol. **83:**593–602.
17. **Goetghebeur, E., J. Liinev, M. Boelaert, and P. Van der Stuyft.** 2000. Diagnostic test analyses in search of their gold standard: latent class analyses with random effects. Stat. Methods Med. Res. **9:**231–248.
18. **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.
19. **Hadgu, A.** 1996. The discrepancy in discrepant analysis. Lancet **348:**592–593.
20. **Hadgu, A.** 1997. Bias in the evaluation of DNA-amplification tests for detecting *Chlamydia Trachomatis*. Stat. Med. **16:**1391–1399.
21. **Hadgu, A., N. Dendukuri, and J. Hilden.** 2005. Evaluation of nucleic acid amplification tests in the absence of a perfect gold-standard test: a review of the statistical and epidemiologic issues. Epidemiology **16:**604–612.
22. **Hawkins, D. M., J. A. Garrett, and B. Stephenson.** 2001. Some issues in resolution of diagnostic tests using an imperfect gold standard. Stat. Med. **20:**1987–2001.
23. **Hui, S. L., and X. H. Zhou.** 1998. Evaluation of diagnostic tests without gold standards. Stat. Methods Med. Res. **7:**354–370.
24. **Jajosky, R. A., P. A. Hall, D. A. Adams, F. J. Dawkins, P. Sharp, W. J. Anderson, J. J. Aponte, G. F. Jones, D. A. Nitschke, C. A. Worsham, N. Adekoya, and T. Doyle.** 2006. Summary of notifiable disease—United States, 2004. Morb. Mortal. Wkly. Rep. **53:**1–79.
25. **Joseph, L., T. W. Gyorkos, and L. Coupal.** 1995. Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. Am. J. Epidemiol. **141:**263–272.
26. **Levy-Guyer, R., and D. E. Koshland.** 1989. The molecule of the year. Science **246:**1543–1544.
27. **Lievano, F. A., M. A. Reynolds, A. L. Waring, J. Ackelsberg, K. M. Bisgard, G. N. Sanden, D. Guris, A. Golaz, D. J. Bopp, R. J. Limberger, and P. F. Smith.** 2002. Issues associated with and recommendations for using PCR to detect outbreaks of pertussis. J. Clin. Microbiol. **40:**2801–2805.
28. **Lind-Brandberg, L., C. Welinder-Olsson, T. Lagergård, J. Taranger, B. Tollfors, and G. Zackrisson.** 1998. Evaluation of PCR for diagnosis of *Bordetella pertussis* and *Bordetella parapertussis* infections. J. Clin. Microbiol. **36:**679–683.
29. **Lipman, H. B., and J. R. Astles.** 1998. Quantifying the bias associated with use of discrepant analysis. Clin. Chem. **1:**108–115.
30. **Loeffelholz, M. J., C. J. Thompson, K. S. Long, and M. J. R. Gilchrist.** 1999. Comparison of PCR, culture, and direct fluorescent-antibody testing for detection of *Bordetella pertussis*. J. Clin. Microbiol. **37:**2872–2876.
31. **McAdam, A. J.** 2000. Discrepant analysis: how can we test a test? J. Clin. Microbiol. **38:**2027–2029.
32. **McCutcheon, A. L.** 2002. Basic concepts and procedures in single- and multiple-group latent class analysis, p. 56–85. *In* J. A. Hagenaars and A. L. McCutcheon (ed.), Applied latent class analysis. Cambridge University Press, Cambridge, United Kingdom.
33. **McNicol, P., S. M. Giercke, M. Gray, D. Martin, B. Brodeur, M. S. Peppler, T. Williams, and G. Hammond.** 1995. Evaluation and validation of a monoclonal immunofluorescent reagent for direct detection of *Bordetella pertussis*. J. Clin. Microbiol. **33:**2868–2871.
34. **Meade, B. D., and A. Bollen.** 1994. Recommendations for use of the polymerase chain reaction in the diagnosis of *Bordetella pertussis* infections. J. Med. Microbiol. **41:**51–55.
35. **Miller, W. C.** 1998. Bias in discrepant analysis: when two wrongs don't make a right. J. Clin. Epidemiol. **51:**219–231.
36. **Miller, W. C.** 1998. Editorial response: can we do better than discrepant analysis for new diagnostic test evaluation? Clin. Infect. Dis. **27:**1186–1193.
37. **Müller, F.-M. C., J. E. Hoppe, and C.-H. Wirsing von König.** 1997. Laboratory diagnosis of pertussis: state of the art in 1997. J. Clin. Microbiol. **35:**2435–2443.
38. **Pepe, M. S.** 2003. The statistical evaluation of medical tests for classification and prediction. Oxford University Press, New York, NY.
39. **Ransohoff, D. F., and A. R. Feinstein.** 1978. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. N. Engl. J. Med. **299:**926–930.
40. **Reed, G. F., F. Lynn, and B. D. Meade.** 2002. Use of coefficient of variation in assessing variability of quantitative assays. Clin. Diagn. Lab. Immunol. **9:**1235–1239.
41. **Riffelmann, M., C. H. Wirsing von König, V. Caro, N. Guiso, et al.** 2005. Nucleic acid amplification tests for diagnosis of *Bordetella* infections. J. Clin. Microbiol. **43:**4925–4929.
42. **Schachter, J.** 1998. Bias in the evaluation of DNA-amplification tests for detecting *Chlamydia trachomatis*. Stat. Med. **17:**1527–1528. (Letter.)
43. **Schachter, J.** 1998. Editorial response: two different worlds we live in. Clin. Infect. Dis. **27:**1181–1185.
44. **Schachter, J.** 2001. In defense of discrepant analysis. J. Clin. Epidemiol. **54:**211–212.
45. **Schachter, J.** 2003. Letter in response to "Statistical guidance on reporting results from studies evaluating diagnostic tests; draft guidance for industry and FDA reviewers" (available at http://www.fda.gov/cdrh/guidance.html). http://www.fda.gov/ohrms/dockets/dailys/03/Sept03/092303/03d-0044-c000007-01-vol1.pdf. Accessed 25 January 2007.

46. **Staquet, M., M. Rozencweig, Y. J. Lee, and F. M. Muggia.** 1981. Methodology for the assessment of new dichotomous diagnostic tests. J. Chronic Dis. **34:**599–610.
47. **Strebel, P. M., S. L. Cochi, K. M. Farizo, B. J. Payne, S. D. Hanauer, and A. L. Baughman.** 1993. Pertussis in Missouri: evaluation of nasopharyngeal culture, direct fluorescent antibody testing, and clinical case definitions in the diagnosis of pertussis. Clin. Infect. Dis. **16:**276–285.
48. **Strebel, P., J. Nordin, K. Edwards, J. Hunt, J. Besser, S. Burns, G. Amundson, A. Baughman, and W. Wattigney.** 2001. Population-based incidence of pertussis among adolescents and adults, Minnesota, 1995–1996. J. Infect. Dis. **183:**1353–1359.
49. **Taranger, J., B. Trollfors, L. Lind, G. Zackrisson, and K. Beling-Holmquist.** 1994. Environmental contamination leading to false-positive polymerase chain reaction for pertussis. Pediatr. Infect. Dis. J. **13:**936–937.
50. **Thibodeau, L. A.** 1981. Evaluating diagnostic tests. Biometrics **37:**801–804.
51. **Torrance-Rynard, V. L., and S. D. Walter.** 1997. Effects of dependent errors in the assessment of diagnostic test performance. Stat. Med. **16:**2157–2175.
52. **U.S. Department of Health and Human Services, Food and Drug Administration, Center for Devices and Radiological Health.** 2003. Statistical guidance on reporting results from studies evaluating diagnostic tests; draft guidance for industry and FDA reviewers. http://www.fda.gov/cdrh/osb/guidance/1428.html.
53. **Valenstein, P. N.** 1990. Evaluating diagnostic tests with imperfect standards. Am. J. Clin. Pathol. **93:**252–258.
54. **van der Zee, A., C. Agterberg, M. Peeters, J. Schellekens, and F. R. Mooi.** 1993. Polymerase chain reaction assay for pertussis: simultaneous detection and discrimination of *Bordetella pertussis* and *Bordetella parapertussis*. J. Clin. Microbiol. **31:**2134–2140.
55. **Vermunt, J. K., and J. Magidson.** 2005. Latent GOLD 4.0 user's guide. Statistical Innovations Inc., Belmont, MA.
56. **Vermunt, J. K., and J. Magidson.** 2005. Technical guide for Latent GOLD 4.0: basic and advanced. Statistical Innovations Inc., Belmont, MA.
57. **Wadowsky, R. M., R. H. Michaels, T. Libert, L. A. Kingsley, and G. D. Ehrlich.** 1996. Multiplex PCR-based assay for detection of *Bordetella pertussis* in nasopharyngeal swab specimens. J. Clin. Microbiol. **34:**2645–2649.
58. **Walter, S. D.** 1999. Estimation of test sensitivity and specificity when disease confirmation is limited to positive results. Epidemiology **10:**67–72.
59. **Walter, S. D., and L. M. Irwig.** 1988. Estimation of test error rates, disease prevalence and relative risk from misclassified data: a review. J. Clin. Epidemiol. **41:**923–937.
60. **Wendelboe, A. M., and A. Van Rie.** 2006. Diagnosis of pertussis: a historical review and recent developments. Expert Rev. Mol. Diagn. **6:**857–864.
61. **Wirsing von König, C. H., S. Halperin, M. Riffelmann, and N. Guiso.** 2002. Pertussis of adults and infants. Lancet Infect. Dis. **2:**744–750.
62. **Zhou, X.-H., N. A. Obuchowski, and D. K. McClish.** 2002. Statistical methods in diagnostic medicine. John Wiley & Sons, Inc., New York, NY.