UW Biostatistics Working Paper Series

10-1-1998

Assessing the Accuracy of a New Diagnostic Test When a Gold Standard Does Not Exist

Todd A. Alonzo
University of Southern California, talonzo@childrensoncologygroup.org

Margaret S. Pepe
University of Washington, mspepe@u.washington.edu

Suggested Citation
Alonzo, Todd A. and Pepe, Margaret S., "Assessing the Accuracy of a New Diagnostic Test When a Gold Standard Does Not Exist" (October 1998). *UW Biostatistics Working Paper Series.* Working Paper 156.
http://biostats.bepress.com/uwbiostat/paper156

This working paper is hosted by The Berkeley Electronic Press (bepress) and may not be commercially reproduced without the permission of the copyright holder.
Copyright © 2011 by the authors

1 Introduction

Medical diagnostic tests play an important role in health care. New diagnostic tests for detecting viral and bacterial infections are continually being developed. The performance of a new diagnostic test is ideally evaluated by comparison with a perfect gold standard test (GS) which assesses infection status with certainty. Many times a perfect gold standard does not exist and a reference test or imperfect gold standard must be used instead. Although statistical methods and techniques are well established for assessing the accuracy of a new diagnostic test when a perfect gold standard exists, such methods are lacking for settings where a gold standard is not available. In this paper we will review some existing approaches to this problem and propose an alternative approach.

To fix ideas we consider a specific example. *Chlamydia trachomatis* is the most common sexually transmitted bacterial pathogen. In women *Chlamydia trachomatis* can result in pelvic inflammatory disease (PID) which can lead to infertility, chronic pelvic pain, and life-threatening ectopic pregnancy. Diagnostic tests for *Chlamydia trachomatis*, among other infections, must be evaluated using specimens from persons whose true infection status cannot be known with certainty. Lacking a perfect GS, cell culture has been used as a reference test. It is generally accepted that culture is nearly 100% specific, but less than 100% sensitive (Black, 1997, Schachter, 1985). The fact that cell culture is not perfect in identifying those with infections causes misclassification of some truly infected subjects as uninfected. Furthermore, this misclassification biases estimates of the accuracy of a new test. If the sensitivity and specificity of culture were known or could be estimated, then existing methods for estimating the accuracy of the new test could be used (Gart & Buck, 1966, Greenberg & Jekel, 1969, Staquet et al., 1981, Baker, 1991). Unfortunately, however, this is not the case.

As technologic advances have been made, more sensitive diagnostic tests for chlamydia have been developed. Tests that use DNA-amplification methods such as ligase chain reaction (LCR) and polymerase chain reaction (PCR) and antigen detection methods such as EIA (enzyme immunoassay) and DFA (direct immunofluorescence) are some of the tests that are thought to be more sensitive than culture. When these tests are compared to cell culture, many of the so-called "false-positive" non-culture test results are in fact true positives or "culture misses". Rather than using culture alone as a reference for evaluating new diagnostic tests, it may be possible to use these more sensitive tests in conjunction with culture as a reference.

Discrepant resolution is one method for assessing the accuracy of new diagnostic tests when a perfect gold standard does not exist but several different imperfect reference tests are available. It has increas-

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

ingly gained popularity, particularly in the detection of infectious diseases. This method will be defined and demonstrated in Section 2 using data from a study in which specimens were tested for *Chlamydia trachomatis*. Recently, concerns with this method have been raised. Some of these will be pointed out in Section 2.3. As an alternative to using discrepant resolution or to using an imperfect gold standard we propose the use of a composite reference standard. This idea shares some appealing features in common with discrepant resolution while avoiding its key problems relating to bias and interpretation. It will be motivated and demonstrated in Section 3. Section 4 will compare analytic expressions for estimates of prevalence and accuracy which are obtained with an imperfect reference test (culture), discrepant resolution, and the composite reference standard. Some cost efficient study designs for evaluating accuracy using the composite reference are considered in Section 5. A method called latent class analysis is becoming popular for the evaluation of clinical diagnostic tests when no gold standard exists and it has recently been suggested as an alternative to discrepant resolution. Its basic approach is very different from the other three methods considered in this paper. Some remarks concerning this technique are given in Section 6. We close with a discussion in Section 7.

2 Discrepant Resolution

2.1 Definition

The goal of discrepant resolution (DR), also known as discrepant analysis, is to obtain accuracy estimates for a new test when the reference test is not a gold standard. Specifically, it uses additional "resolver" tests to resolve the discrepant results between the new diagnostic test and the imperfect reference test. For the purpose of discussion the imperfect reference will be referred to as a culture test in this paper. In addition to its use with tests for detecting *Chlamydia trachomatis* (Polaneczky et al., 1998, Crotchfelt et al., 1998, Gaydos et al., 1998), discrepant resolution has been used to study the performance of tests for *Neisseria gonorrhoeae* (Ciemins et al., 1997, Young et al., 1997, Ching et al., 1995), *Closttridium difficile* (Schue et al., 1994, DeGirolami et al., 1992), *Mycobacterium tuberculosis* (Bergmann & Woods, 1997, Gamboa et al., 1997, Smith et al., 1997), *Toxoplasma gondii* (Crouch, 1995), *Helicobacter pylori* (Graham et al., 1996, Pronovost et al., 1994), *pnuemocystis carinii* (Mathis et al., 1997), hepatitis C virus (Kessler et al., 1997, Morris et al., 1996), cytomegalovirus (Roseff & Campos, 1993, Zweygberg et al., 1990), and herpes simplex virus (Cullen et al., 1997, Dascal et al., 1989).

There are two stages to the discrepant resolution algorithm. In the first stage all n specimens are

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

tested with the new test and with culture. The results of the testing in stage 1 can be summarized in a contingency table (stage 1 of Table 1). The number in a specific cell of the contingency table represents the number of specimens with that specific combination of test results. The first and second subscripts denote results of the new test and culture, respectively. For example, n_{+-} is the number of specimens which are considered new test positive and culture negative.

Culture based estimates of the accuracy of the new test can be calculated using the information obtained in stage 1 of Table 1. The prevalence of infection, sensitivity, specificity, positive predicted value (PPV), and negative predicted value (NPV) of the new test when culture is used as the standard can be estimated as follows:

$$\begin{align}
\text{prevalence}_C &= \frac{n_{++} + n_{-+}}{n} \tag{1}\\
\text{sensitivity}_C &= \frac{n_{++}}{n_{++} + n_{-+}} \tag{2}\\
\text{specificity}_C &= \frac{n_{--}}{n_{--} + n_{+-}} \tag{3}\\
\text{PPV}_C &= \frac{n_{++}}{n_{++} + n_{+-}} \tag{4}\\
\text{NPV}_C &= \frac{n_{--}}{n_{--} + n_{-+}} \tag{5}
\end{align}$$

Since culture is known to be an imperfect reference test, DR attempts to improve the reference by retesting those specimens for which the two tests disagree (n_{+-} and n_{-+}) in the first stage using a resolver test. Results from the resolver test are used to update the contingency table formed from stage 1 testing and result in a new contingency table (stage 2 of Table 1). Only specimens which were re-tested with the resolver have a third subscript indicating the resolver test result. For example, specimens that are new test positive/culture negative and positive by the resolver are considered positive by discrepant resolution and are denoted n_{+-+}. When the discrepant resolution algorithm is used, the following estimates of prevalence, sensitivity, specificity, PPV, and NPV are obtained:

$$\begin{align}
\text{prevalence}_{DR} &= \frac{n_{++} + n_{+-+} + n_{-++}}{n} \tag{6}\\
\text{sensitivity}_{DR} &= \frac{n_{++} + n_{+-+}}{n_{++} + n_{+-+} + n_{-++}} \tag{7}\\
\text{specificity}_{DR} &= \frac{n_{--} + n_{-+-}}{n_{--} + n_{-+-} + n_{+--}} \tag{8}\\
\text{PPV}_{DR} &= \frac{n_{++} + n_{+-+}}{n_{++} + n_{+-+} + n_{+--}} \tag{9}\\
\text{NPV}_{DR} &= \frac{n_{--} + n_{-+-}}{n_{--} + n_{-+-} + n_{-++}} \tag{10}
\end{align}$$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

2.2 Example

Wu et al. (1991) conducted a study in which specimens from 324 men and women attending two STD clinics in China and Taiwan were tested for *Chlamydia trachomatis* using cell culture, PCR, and EIA. In this study the researchers wanted to assess the accuracy of EIA. The results from cell culture and EIA are summarized in stage 1 of Table 2. If culture is considered the reference, then using equation (1) suggests the prevalence of *Chlamydia trachomatis* is 23/324=0.071. Furthermore, based on equations (2)-(5), when culture is used as the standard the estimated sensitivity of EIA is 20/23=0.870, estimated specificity is 294/301=0.977, estimated PPV is 20/27=0.741, and estimated NPV is 294/297=0.990. In stage 2 of discrepant resolution, the 7 specimens which were positive by EIA and negative by culture and the 3 culture positive/EIA negative specimens are re-tested with the resolver, PCR. Four of the 7 EIA positive/culture negative specimens were considered positive by PCR and 1 of the 3 EIA negative/culture positives was PCR negative. The results after resolving the discrepancies are summarized in stage 2 of Table 2. Applying equations (6)-(10) suggest the prevalence of infection is 26/324=0.080, sensitivity of EIA is (20+4)/(20+4+3-1)=24/26=0.923 and specificity is (294+1)/(294+1+7-4)=295/298=0.990. Furthermore, the PPV and NPV of EIA are estimated to be 24/27=0.889 and 295/297=0.993, respectively.

2.3 Concerns with DR

Although DR has been in use since at least 1984, the first concerns about it were raised only recently (Hadgu, 1996). A key problem with DR has to do with the ambiguous interpretation of the sensitivity and specificity estimates obtained. The standard, relative to which accuracy is measured, depends intrinsically on the results of the new test, as can be seen by noting that the denominators of (7) and (8) involve the result of the new test. When evaluating accuracy of a new test, it is clearly imperative that it be compared to a standard which is independent of the new test itself. DR violates this principle.

Biases in DR estimates of accuracy have been described. Miller (1998), Hadgu (1997), and Lipman & Astles (1998) algebraically and numerically demonstrated the bias in sensitivity and specificity estimates in the ideal setting where the resolver test is a perfect gold standard. They showed that sensitivity and specificity estimates obtained from DR are biased upwards, so that the new test appears to be more accurate than it really is. Green et al. (1998) showed that when the resolver test is not a perfect gold standard even larger biases are possible and in some situations biases in DR estimates are smaller than the biases resulting from culture (an imperfect reference test). Of particular concern is the setting where

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

errors made by the resolver are similar in nature to those made by the new test.

3 Composite Reference Standards

3.1 Definition

We propose that the results of several imperfect reference tests can be used in combination to define a standard against which a new test can be compared. The tests in the combination cannot, of course, include the new test itself. We call the resultant standard a composite reference standard (CRS). In the chlamydia setting, for example, one might consider that any specimen that is culture positive or resolver test positive is CRS positive and any specimen that is culture negative and resolver negative is CRS negative independent of the results of the new test. In this setting, both the culture and resolver tests are likely to be highly specific so that the CRS should also be highly specific while more sensitive than either test alone. Other combinations of culture and resolver test results could be used to define a CRS that may be relevant in other settings. The resolver could even be a repeat application of culture testing.

The use of a CRS shares some of the appealing features of the discrepant resolution approach while avoiding its most problematic disadvantages. Like discrepant resolution, the CRS approach allows one, (i) to use several sources of information in order to assess if an infection is present; and (ii) to ascertain the reference test information in a sequential fashion which avoids the need for redundant testing. In the chlamydia setting with the aforementioned CRS, specimens that test positive with culture at the first stage would not need to be tested with a resolver at the second stage, since they are already known to be CRS positive by definition. In contrast to discrepant resolution however, the CRS reference does not depend on the results of the new test under investigation. One drawback of the CRS is that it requires testing the typically large number of specimens negative by both culture and the new test. However, a solution to this problem is proposed in Section 5.

The particular CRS which we will pursue in detail here is the aforementioned CRS which is defined as being positive if either reference test, culture or a resolver, is positive and negative otherwise. This definition was used by Jang et al. (1992) for the chlamydia setting. To assess the performance of the new test using CRS, there are two stages to testing (Table 3). As with DR, in the first stage all specimens are tested by the new test and culture. At the second stage only those specimens which are culture negative at the first stage are tested with the resolver since by definition all culture positives are CRS positive. Notation for the results of both stages of testing is displayed in Table 3. For example, n_{+-+} is the number

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

of new test positive/culture negative/resolver positive specimens. The results from both stages of testing are combined to get valid estimates of the performance of the new test using CRS.

It can be shown that based on these data the maximum likelihood estimator of sensitivity is

$$\text{sensitivity}_{CRS} = \frac{n_{++} \ + \ n_{+-+}}{n_{++} \ + \ n_{+-+} \ + \ n_{-+} \ + \ n_{--+}} \tag{11}$$

and the maximum likelihood estimate of the specificity of the new test using the CRS is

$$\text{specificity}_{CRS} = \frac{n_{---}}{n_{---} \ + \ n_{+--}} \tag{12}$$

Similarly

$$\begin{aligned} \text{prevalence}_{CRS} &= \frac{n_{++} \ + \ n_{+-+} \ + \ n_{-+} \ + \ n_{--+}}{n} \quad (13) \\ \text{PPV}_{CRS} &= \frac{n_{++} \ + \ n_{+-+}}{n_{++} \ + \ n_{+-+} \ + \ n_{+--}} \quad (14) \\ \text{NPV}_{CRS} &= \frac{n_{---}}{n_{---} \ + \ n_{-+} \ + \ n_{--+}} \quad (15) \end{aligned}$$

3.2 Example

Again consider the data from the study done by Wu et al. (1991). Contingency tables summarizing the two stages of CRS for these data are given in Table 4. The first stage in assessing the performance of EIA using the CRS is the same as that in the DR algorithm. In the second stage the 301 culture negative specimens are tested with PCR. Two of the 294 EIA negative/culture negative specimens were positive by PCR, and as noted above 4 of the 7 EIA positive/culture negatives were PCR positive. Therefore, an additional 6 specimens are considered to be infected when the CRS is used as the standard. Using equations (11)-(15), the maximum likelihood estimate of prevalence is 29/324=0.090, of sensitivity is (20+4)/(20+4+3+2)=24/29=0.828, of specificity is (294-2)/(294-2+7-4)=292/295=0.990, of PPV is 24/27=0.889, and of NPV is 292/297=0.983.

3.3 Infection Missed by Culture and the New Test

In typical studies of low prevalence cohorts, a majority of the specimens will be negative at the first stage according to both culture and the new test. A major problem with the DR algorithm is that it does not test the large number of new test negative/culture negative specimens with the resolver. The CRS method, on the other hand, does test this group of specimens. In Section 5.2 we will examine, in the general setting, the importance of testing these specimens with the resolver test. To gain some insight into this question

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

in a real study, we determined (Figure 1) the effects of varying values of n_{--+} on the CRS estimates of sensitivity and specificity in the context of the data from Wu et al. (1991).

Figure 1(a) suggests the estimated sensitivity decreases sharply as n_{--+} increases. If none of the culture negative/new test negative specimens tested positive with PCR, the resolver, then the sensitivity would be 0.889. However, if for example, 9 (3.06%) tested positive with the resolver, the estimate of sensitivity drops substantially to 0.667. Thus the sensitivity estimate depends heavily on the results of resolver testing of the new test negative/culture negative specimens. The specificity estimate on the other hand does not as can be seen from Figure 1(b). Estimated specificity decreases slowly with n_{--+}. If none of the new test negative/culture negative specimens test positive with PCR, the specificity is 0.99, while the specificity only drops to 0.98 when 147 (50%) of these specimens test positive with PCR. In practice such a high percentage of positives would be extremely unlikely.

4 Comparison of Accuracy Parameters

Thus far we have described three reference standards for assessing the performance of a new diagnostic test, an imperfect gold standard (culture), discrepant resolution, and a composite reference standard. Equations for estimating prevalence of infection, sensitivity, specificity, PPV, and NPV of the new test using each of the three methods have been presented. Therefore, analytic comparisons of the estimates obtained from each method can be made. The results of such comparisons are summarized in Table 5.

Comparing equations (1) and (6) suggests that the prevalence of infection when the DR algorithm is used can be lower or higher than the prevalence estimated using culture as the standard with equality when n_{-+-} equals n_{+-+}. On the other hand, the sensitivity, specificity, PPV, and NPV of the new test when DR is used is always larger than or equal to the estimates obtained using culture alone as the standard. This is not surprising since resolving discrepant results can only increase the number of specimens in the diagonal cells of the contingency table.

The CRS was defined so that not only culture positive specimens, but also resolver positive specimens were considered to be infected. Therefore, it is not surprising that prevalence of infection when the CRS is used is always the same or higher than the prevalence when culture alone is used as the standard. Furthermore, since the prevalence estimated using the CRS is always greater than or equal to the culture based estimate of prevalence, it makes sense that the PPV estimated using the CRS is larger than or equal to the culture based estimate of PPV, and the NPV estimated using the CRS is less than or equal to the

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

culture based estimate of NPV. Comparisons of sensitivity and specificity are not as straightforward.

The estimate of sensitivity when CRS is used is a weighted average of the culture based estimate of sensitivity and $n_{+-+}/(n_{+-+} + n_{--+})$. The latter is the proportion of the resolver positive/culture negative specimens that were new test positive, i.e., the sensitivity of the new test to samples testing positive only to the resolver and not to culture. Since the CRS based estimate of the sensitivity of the new test is a weighted average, the following statement is true.

A. $\text{sensitivity}_C \geq \text{sensitivity}_{CRS}$, if and only if $\text{sensitivity}_C \geq \frac{n_{+-+}}{n_{+-+} + n_{--+}}$.

A similar statement can be made in regards to specificity comparisons upon noting that the estimate of the specificity of the new test using the CRS is a weighted average of the culture based estimate of specificity and $n_{--+}/(n_{+-+} + n_{--+})$.

B. $\text{specificity}_C \leq \text{specificity}_{CRS}$, if and only if $\text{specificity}_C \geq \frac{n_{--+}}{n_{+-+} + n_{--+}}$.

Observe that in statements A and B equality on the left implies and is implied by equality on the right.

General comparisons of the estimates obtained using the CRS and the DR algorithm can also be made. The prevalence of infection when infection is defined using the CRS is always greater than or equal to the prevalence when DR is used. On the other hand, comparing the equations for CRS and DR we see that sensitivity, specificity, and NPV estimates resulting from the CRS are less than or equal to the corresponding DR estimates. Interestingly, equations (9) and (14) are the same, so that the PPV estimates resulting from the two methods are the same.

In Table 6, using data from Wu et al. (1991), estimated accuracy and prevalence parameters calculated using culture, DR, and CRS are displayed. The inequalities in Table 5 are borne out with the real data. The estimate of sensitivity using CRS appears to be less than the culture based estimate, whereas specificity appears to be better when the CRS is used. Of the 6 specimens testing negative with culture and positive with the resolver only 4 test positive with the new test. Thus the right hand side of statements A and B held, implying the left hand side inequalities.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

5 CRS with Modified Sampling

One drawback to using the CRS in practice is that typically there are a large number of new test negative/culture negative specimens and testing all of these with the resolver test can become expensive and time consuming. Indeed a motivation for the DR approach is that none of these need to be re-tested with the resolver. A strategy to reduce cost and time, while employing the CRS is to select subsets of the new test negative/culture negative specimens and of the new test positive/culture negative specimens for testing with the resolver. We now show how estimates of accuracy parameters associated with the new test using the CRS can be calculated based on data from such a study design.

5.1 Modified Accuracy Estimates

Let m_{--} be the number of new test negative/culture negative specimens selected from the n_{--} available, that are to be tested with the resolver. Then m_{--}/n_{--} is the fraction of the new test negative/culture negative specimens selected. Let m_{--+} and m_{---} denote the number of these specimens that are resolver positive and resolver negative, respectively. Similarly, m_{+-+} denotes the number of specimens from the subset of size m_{+-} that are resolver positive while m_{+--} corresponds to the number of such specimens that are resolver negative. The multinomially based likelihood for these data is:

$$c_0\, {p_{++}}^{n_{++}}\ \ {p_{-+}}^{n_{-+}}\ \ {p_{+-}}^{(n_{+-}-m_{+-})}\ \ {p_{--}}^{(n_{--}-m_{--})}\ \ {p_{+-+}}^{m_{+-+}}\ \ {p_{+--}}^{m_{+--}}\ \ {p_{--+}}^{m_{--+}}\ \ {p_{---}}^{m_{---}}$$

where c_0 is a normalizing constant and cell probabilities are denoted by p with analogous subscript notation for the outcomes of the new test, culture and resolver tests.

Solving the likelihood equations yields the following estimates:

$$\begin{aligned}
\hat{p}_{++} &= \frac{n_{++}}{n};\ \hat{p}_{-+} = \frac{n_{-+}}{n} \\
\hat{p}_{+-+} &= \frac{m_{+-+}(n_{+-}/m_{+-})}{n};\ \hat{p}_{+--} = \frac{m_{+--}(n_{+-}/m_{+-})}{n} \\
\hat{p}_{--+} &= \frac{m_{--+}(n_{--}/m_{--})}{n};\ \hat{p}_{---} = \frac{m_{---}(n_{--}/m_{--})}{n}
\end{aligned}$$

Since m_{+-}/n_{+-} and m_{--}/n_{--} are the sampling fractions for the second stage, the estimated cell frequencies associated with the resolver test results if all specimens were re-tested (i.e. the numerators of $\hat{p}_{+-+}, \hat{p}_{+--}, \hat{p}_{--+}$ and $\hat{p}_{---}$) are the observed cell frequencies multiplied by the inverse of the sampling fraction. Inserting these estimates into expressions for prevalence and accuracy yields the following:

$$\text{prevalence}_{CRS} = \frac{n_{++} + m_{+-+}(n_{+-}/m_{+-}) + n_{-+} + m_{--+}(n_{--}/m_{--})}{n}$$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

$$\begin{aligned}
\text{sensitivity}_{CRS} &= \frac{n_{++} + m_{+-+}(n_{+-}/m_{+-})}{n_{++} + m_{+-+}(n_{+-}/m_{+-} + n_{-+} + m_{--+}(n_{--}/m_{--}))} \\
\text{specificity}_{CRS} &= \frac{m_{---}(n_{--}/m_{--})}{m_{---}(n_{--}/m_{--}) + m_{+--}(n_{+-}/m_{+-})} \\
\text{PPV}_{CRS} &= \frac{n_{++} + m_{+-+}(n_{+-}/m_{+-})}{n_{++} + m_{+-+}(n_{+-}/m_{+-}) + m_{+--}(n_{+-}/m_{+-})} \\
\text{NPV}_{CRS} &= \frac{m_{---}(n_{--}/m_{--})}{m_{---}(n_{--}/m_{--}) + n_{-+} + m_{--+}(n_{--}/m_{--})}
\end{aligned}$$

5.2 Choice of Sampling Fractions

Expressions for the variances of the modified estimates of sensitivity and specificity are provided in the Appendix. Standard errors can be calculated by substituting estimates for the unknown components as described there. Sample sizes and desirable sampling fractions can be based on these expressions. Confidence intervals for sensitivity and specificity can also be based on these standard errors in large samples. A better strategy for confidence intervals is perhaps to use a logit transformation of the estimators and the delta method for the standard error of the logit transform in order to calculate confidence limits for the logit transform. Confidence limits for parameters then are calculated as logistic functions of these limits. Newcombe (1998) compares different methods for calculating confidence intervals for a proportion including using a logit transformation.

The variances of the parameter estimates depend on the proportions of specimens chosen for testing in the second stage. To understand how these sampling fractions influence variability in the specificity and sensitivity estimates we calculated asymptotic variance expressions with various sampling fractions using the cell probabilities from Wu's data. Figure 2 presents the ratio of the variance when all n_{--} specimens testing negative with culture and EIA are tested with resolver, to the variance when a fraction m_{--}/n_{--} are tested with resolver, i.e. the asymptotic relative efficiency (ARE). In these plots the sampling fraction, m_{+-}/n_{+-}, is set to 1, i.e., all specimens testing positive with EIA and negative with the culture are tested with the PCR resolver. It appears that the sampling fraction in the (-,-) cell has little effect on the precision with which the specificity is estimated. Additional numerical work (not shown) suggests that this will be the case in most practical settings. However, the fraction of EIA negative/culture negative specimens tested with the resolver does have a substantial effect on the precision of the sensitivity estimate. For example, with Wu's data the standard error of the sensitivity estimate is 15% larger when 50% of the culture negative/EIA negative specimens are re-tested than when all n_{--} specimens are re-tested with

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

the resolver.

Figure 3 shows how the asymptotic relative efficiencies of the estimates vary with the sampling fraction m_{+-}/n_{+-}; i.e., when a fraction of the culture negative/EIA positive specimens are selected for stage 2 testing. The precision of the sensitivity estimate is unaffected but that of the specificity is strongly influenced by m_{+-}/n_{+-}. In practice however one will most likely test all of the culture negative/new test positive specimens because there will be relatively few of them and because they will be of scientific or clinical interest.

6 Latent Class Analysis

6.1 Definition and Concerns

A statistical solution to the imperfect gold standard problem put forth by Walter & Irwig (1988) amongst others is latent class analysis. The basic idea of latent class analysis is to assume that there exists some unobservable infection status and to relate the observed diagnostic test results to it with a statistical model. It requires that a minimum of three (imperfect) diagnostic tests be measured on every specimen, in contrast to DR or to the modified sampling CRS approach. Maximum likelihood techniques then yield estimates of infection prevalence and of test accuracy for each of the tests.

There are at least three problems with this approach. First, infection is not an explicitly defined entity in this approach. It is the intangible which links the results of the observed diagnostic tests. We prefer the CRS approach over the latent class approach in part because the composite reference is well-defined in the CRS method. A second problem with latent class analysis is that it relies on a statistical model which cannot be fully tested. With three diagnostic tests for example, the critical assumption is that test results are statistically independent given infection status. It is impossible to examine the validity of this assumption when only three diagnostic tests have been applied. More complex models can be applied when more tests are available but unverifiable assumptions about dependence are still required. Finally, the estimates calculated using this approach are not simple explicit functions of the data. For the clinician, the estimates are output from a “black-box” statistical algorithm. For the statistician, it is hard to get an intuitive understanding for how the data affect the estimates. In the next section we derive some analytic expressions which may be helpful in this regard.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

6.2 Parameter Estimates

Let p denote infection prevalence, ϕ_k denote the sensitivity of the k^{th} test and θ_k its specificity. We suppose that there are three diagnostic tests and hence that the conditional independence statistical model is used. With Y_{jk} denoting the result of the k^{th} test on the j^{th} specimen, the likelihood for the three test results from the j^{th} specimen under the conditional independence model is

$$L_j = p \prod_k \phi_k^{Y_{jk}} (1-\phi_k)^{1-Y_{jk}} + (1-p) \prod_k \theta_k^{1-Y_{jk}} (1-\theta_k)^{Y_{jk}}$$

Differentiating the log-likelihood, $\sum_j \log(L_j)$, with respect to p, ϕ_k and θ_k $k = 1, 2, 3$ and setting the derivatives to zero yields the following equations:

$$\begin{aligned} p &= \sum w_j / n & (16) \\ \phi_k &= \sum w_j Y_{jk} / \sum w_j & (17) \\ \theta_k &= \sum (1-w_j)(1-Y_{jk}) / \sum (1-w_j) & (18) \end{aligned}$$

where $w_j = a_j/(a_j + b_j), a_j = p \prod_k \phi_k^{Y_{jk}} (1-\phi_k)^{1-Y_{jk}}$ and $b_j = (1-p) \prod_k \theta_k^{1-Y_{jk}} (1-\theta_k)^{Y_{jk}}$. The weights, w_j, can be interpreted as conditional probabilities calculated under the conditional independence model

$$w_j = P(D_j = 1 | Y_{j1}, Y_{j2}, Y_{j3}).$$

Note that (16)–(18) do not give explicit expressions for the parameter estimates as functions of the data since the weights are themselves functions of the parameters. Rather, at the maximum likelihood solutions for the parameters $\{p, (\phi_k, \theta_k)\ k = 1, 2, 3\}$, the result is that these equations are satisfied. Thus the prevalence estimate is an average of the estimated probabilities, $\hat{p} = \sum \hat{P}(D_j = 1 | Y_{j1}, Y_{j2}, Y_{j3})/n$. The sensitivity (1-specificity) estimate for the k^{th} test is a weighted average of the k^{th} test results weighted by the estimated probability of that observation being from a specimen with (without) an infection.

6.3 Conditional Dependence

The above formulation now allows us to address the question of how violation of the conditional independence assumption can affect parameter estimation. Observe that if the weights w_j were calculated under the true model, then (16)–(18) should be satisfied asymptotically. Comparing weights calculated under correct and incorrect models can give insight into biases which can result under violation of the conditional independence assumption.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Consider for example that the three diagnostic tests for infection are 100% specific but only 80% sensitive and that the true prevalence of infection is 50%, say. For the sake of illustration suppose that the tests are highly dependent with $P(Y_2 = 1, Y_3 = 1|Y_1 = 1, D = 1) = 1$ and $P(Y_2 = 0, Y_3 = 0|Y_1 = 0, D = 1) = 1$. That is, tests 2 and 3 identify the same specimens as having an infection as does test 1 and hence are conditionally dependent. Only two types of triples (Y_1, Y_2, Y_3) occur in the data, (1,1,1) and (0,0,0), with frequencies of 40% and 60%, respectively. Observe that the true conditional probabilities are: $P(D = 1|(0, 0, 0)) = P((0, 0, 0)|D = 1)P(D = 1)/P(0, 0, 0) = 0.2 \times 0.5/0.6 = 0.167$ and $P(D = 1|(1, 1, 1)) = 1$. Under the conditional independence model $P(D = 1|(1, 1, 1))$ remains at 1 but $P(D = 1|(0, 0, 0))$ is reduced substantially to $P(D = 1|(0, 0, 0)) = 0.0079$ from its true value of 0.167.

From equation (16) we see that the latent class analysis which assumes conditional independence therefore yields a prevalence estimate below the true prevalence. In equation (17) the sensitivity estimate gives very little weight to negative test results yielding a sensitivity estimate which is biased too high. Similarly, the specificity estimate will overweight negative test results yielding a high estimate of specificity. These considerations suggest that when test results are positively correlated for infected and/or uninfected specimens, latent class analysis will yield accuracy estimates which are too high. This result is corroborated by simulation results of Torrance-Rynard & Walter (1997).

6.4 Example

The latent class method was applied to data from Wu et al. (1991). This method estimated the prevalence of Chlamydia to be 0.081 and yielded the estimates of sensitivity and specificity shown in Table 7 for each of the three tests, EIA, culture, and PCR. Comparing these results to those given in Table 6 suggests that the prevalence estimate given by latent class analysis is in between the estimates obtained when culture is considered the reference and when the CRS is used. The estimate of sensitivity for EIA resulting from latent class analysis is higher than that obtained using culture or the CRS as the reference but not as high as that using DR. All of the approaches including latent class analysis indicate that EIA is highly specific. Latent class analysis suggests that culture and PCR also have near perfect specificities, 0.997 and 0.995 respectively.

Latent class analysis assumes that the three tests are conditionally independent. Since the three tests are based on different clinical methods, antigen detection, cell culture, and DNA-amplification, the tests are less likely to make the same type of errors than if, for example, two of the three tests were DNA-amplification tests. However, conditional dependence may still exist. For example, it is possible that the

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

tests misclassify more specimens from persons with a borderline case of chlamydia than those specimens from persons with a severe case of chlamydia. In this case the conditional independence assumption would be violated.

7 Discussion

In this paper we have reviewed several approaches for assessing the performance of a new diagnostic test when a gold standard does not exist. Deficiencies associated with use of the imperfect reference, discrepant resolution, and latent class analysis were identified. We have proposed that in some settings an alternative approach, called the CRS, may be employed. It combines the results of several imperfect reference tests to define a better standard against which a new test can be compared. Some of the advantages of this method are that it allows one to use several sources of information in order to assess if an infection is present, and to ascertain the reference test information in a sequential fashion which avoids the need for redundant testing. Most importantly, it is well-defined based on observable quantities, and the reference test is not affected by results of the new test under investigation. Thus, the results and standard against which the new test is compared are easy to interpret.

Despite the fact that concerns with DR have been raised and substantial biases have been quantified, this algorithm is still in use because some researchers believe accepting the problems with DR is better than the alternative of using culture which is known to be an imperfect GS. The CRS is likely more palatable than culture alone and thus is a viable alternative to DR. As it turns out, several different CRS have already been used to study the performance of new tests to detect *Chlamydia trachomatis* along with other diseases such as *Bordetella pertussis*. Some researchers have used a CRS defined just as we have (Jang et al., 1992, McNicol et al., 1995, Sellors et al., 1991). Others have defined the CRS as a combination of results from multiple resolver tests so that specimens positive by culture or at least one of the resolvers are considered to be CRS positive (Chernesky et al., 1990). A contribution of this paper is to formalize the notion of a CRS and to suggest how accuracy can be estimated statistically relative to a CRS using sequential sampling. We propose this as a valid alternative to discrepant resolution, as a constructive step in response to criticism of the discrepant resolution approach.

An attribute of the CRS method is that it does not require that all specimens be tested with the resolver reference test. With the specific CRS we have studied in this paper only culture negative specimens require re-testing and of them only fractions of the culture negative/new test negative and culture negative/new

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

test positive specimens need be re-tested. The key specimens are the new test negative/culture negative ones since there are likely to be many of these and they do not need re-testing with the DR approach. Green et al. (1998) have shown that discrepant resolution estimates specificity reasonably well but does not estimate sensitivity well. Our results also show that inference about sensitivity but not specificity depends on re-testing of these specimens with a resolver. By using the CRS method, we provide a well-defined (in contrast to latent class analysis) and valid (in contrast to DR) estimate of sensitivity. However, the precision of the sensitivity estimate we obtain is reduced by omitting some of the new test negative/culture negative tests from re-testing.

Although the estimates obtained using the discrepant resolution algorithm usually overestimate the sensitivity and NPV calculated using the CRS, the DR estimate of PPV is identical to that obtained using the CRS and the DR estimate of specificity is reasonably close to the CRS estimate. Therefore, if the sole purpose of a study is to determine the PPV or specificity of a new diagnostic test, then the DR algorithm which re-tests a much smaller number of specimens with the resolver than the CRS method could be used. However, if the DR algorithm is used to obtain these estimates, then care must be taken when interpretating the results.

As technologic advances are made, new diagnostic tests for viral and bacterial infections will be developed. Advances in the statistical methods available to assess the accuracy of these tests must also be made, but not at the expense of interpretation or ease of implementation. We propose the composite reference standard as one such method when a gold standard reference test does not exist. In this paper we have assumed that test results from the new test are binary. We are currently developing methods that allow one to assess the accuracy of new tests that yield continuous results.

ACKNOWLEDGEMENTS

This research was supported by NIH grant R01 GM54438.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

References

Baker, S. (1991). Evaluating a new test using a reference test with estimated sensitivity and specificity. *Communication in Statistics: Theory and Methods* **20**, 2739–2752.

Bergmann, J. & Woods, G. (1997). Mycobacterial growth indicator tube for susceptibility testing of *Mycobacterium tuberculosis* to isoniazid and rifampin. *Diagnostic Microbiology and Infectious Disease* **28**, 153–156.

Black, C. (1997). Current methods of laboratory diagnosis of *chlamydia trachomatis* infections. *Clinical Microbiology Reviews* **10**, 160–184.

Chernesky, M., Castriciano, S., Sellors, J., Stewart, I., Cunningham, I., Landis, S., Seidelman, W., Grant, L., Devlin, C., & Mahony, J. (1990). Detection of *Chlamydia trachomatis* antigens in urine as an alternative to swabs and cultures. *Journal of Infectious Diseases* **161**, 124–126.

Ching, S., Lee, H., Hook, E., Jacobs, M., & Zenilman, J. (1995). Ligase chain reaction for detection of *neisseria gonorrhoeae* in urogenital swabs. *Journal of Clinical Microbiology* **33**, 3111–3114.

Ciemins, E., Borenstein, L., Dyer, I., Cordero, E., Courtney, J., Harvey, S., & Richwald, G. (1997). Comparisons of cost and accuracy of DNA probe test and culture for the detection of *Neisseria gonorrhoeae* in patients attending public sexually transmitted disease clinics in los angeles county. *Sexually Transmitted Diseases* **24**, 422–428.

Crotchfelt, K., Pare, B., Gaydos, C., & Quinn, T. (1998). Detection of *Chlamydia Trachomatis* assay (AMP CT) in urine specimens from men and women and endocervical specimens from women. *Journal of Clinical Microbiology* **36**, 391–394.

Crouch, C. (1995). Enzyme immunoassays for IgG and IgM antibodies to *Toxoplasma gondii* based on enhanced chemiluminescence. *Journal of Clinical Pathology* **48**, 652–657.

Cullen, A., Long, C., & Lorincz, A. (1997). Rapid detection and typing of herpes simplex virus DNA in clinical specimens by the hybrid capture ii signal amplification probe test. *Journal of Clinical Microbiology* **35**, 2275–2278.

Dascal, A., Chan-Thim, J., Morahan, M., Portnoy, J., & Mendelson, J. (1989). Diagnosis of herpes simplex virus infection in a clinical setting by a direct antigen detection enzyme immunoassay kit. *Journal of Clinical Microbiology* **27**, 700–704.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

DeGirolami, P., Hanff, P., Eichelberger, K., Longhi, L., Teresa, H., Pratt, J., Cheng, A., Letournea, J., & Thorne, G. (1992). Multicenter evaluation of a new enzyme immunoassay for detection of *Clostridium difficile* enterotoxin A. *Journal of Clinical Microbiology* **30**, 1085–1088.

Gamboa, F., Manterola, J., Lonca, J., Vinado, B., L. Matas, M. G., Manzano, J., Rodrigo, C., Cardona, P., Padilla, E., Dominguez, J., & Ausina, V. (1997). Rapid detection of *Mycobacterium tuberculosis* in respiratory specimens, blood and other non-respiratory specimens by amplification of rRNA. *International Journal of Tuberculosis and Lung Disease* **1**, 542–555.

Gart, J. & Buck, A. (1966). Comparison of a screening test and a reference test in epidemiologic studies. ii: A probabilistic model for the comparison of diagnostic tests. *American Journal of Epidemiology* **83**, 593–602.

Gaydos, C., Crotchfelt, K., Howell, M., Kralian, S., Hauptman, P., & Quinn, T. (1998). Molecular amplification assays to detect chlamydial infections in urine specimens from high school female students and to monitor the persistence of chlamydial DNA after therapy. *Journal of Infectious Diseases* **177**, 417–424.

Graham, D., Evans, D., Peacock, J., Baker, J., & Schrier, W. (1996). Comparison of rapid serological tests (FlexSure HP and QuickVue) with conventional ELISA for detection of *Helicobacter pylori* infection. *American Journal of Gastroenterology* **91**, 942–948.

Green, T., Black, C., & Johnson, R. (1998). Evaluation of bias in diagnostic-test sensitivity and specificity estimates computed by discrepant analysis. *Journal of Clinical Microbiology* **36**, 375–381.

Greenberg, R. & Jekel, J. (1969). Some problems in the determination of false negative rates of tuberculin tests. *American Review of Respiratory Disease* **100**, 645.

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*. *Statistics in Medicine* **16**, 1391–1399.

Jang, D., Sellors, J., Mahony, J., Pickard, L., & Chernesky, M. (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. *Sexually Transmitted Diseases* **19**, 315–319.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

Kessler, H., Dragon, E., Pierer, K., Santner, B., Liao, Y., Stunzner, D., Stelzl, E., & Marth, E. (1997). Performance of the automated COBAS AMPLICOR system for the detection of hepatitis C virus RNA. *Clinical and Diagnostic Virology* **7**, 139–145.

Lipman, H. & Astles, J. (1998). Quantifying the bias associated with use of discrepant analysis. *Clinical Chemistry* **44**, 108–115.

Mathis, A., Weber, R., Kuster, H., & Speich, R. (1997). Simplified sample processing combined with a sensitive one-tube nested PCR assay for detection of *Pneuomocystis carinii* in respiratory specimens. *Journal of Clinical Microbiology* **35**, 1691–1695.

McNicol, P., Giercke, S., Gray, M., Martin, D., Brodeur, B., Peppler, M., Williams, T., & Hammond, G. (1995). Evaluation and validation of a monoclonal immunofluorescent reagent for direct detection of *Bordetella pertussis*. *Journal of Clinical Microbiology* **33**, 2868–2871.

Miller, W. (1998). Bias in discrepant analysis: when two wrongs don't make a right. *Journal of Clinical Epidemiology* **51**, 219–231.

Morris, T., Robertson, B., & Gallagher, M. (1996). Rapid reverse transcription-PCR detection of hepatitis C virus RNA in serum using the TaqMan flurogenic detection system. *Journal of Clinical Microbiology* **34**, 2933–2936.

Newcombe, R. (1998). Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine* **17**, 857–872.

Polaneczky, M., Quigley, C., Pollock, L., Dulko, D., & Witkins, S. (1998). Use of self-collected vaginal specimens for detection of *Chlamydia trachomatis* infection. *Obstetrics and Gynecology* **91**, 375–378.

Pronovost, A., Rose, S., Pawlak, J., Robin, H., & Schneider, R. (1994). Evaluation of a new immunodiagnostic assay for *Helicobacter pylori* antibody detection: Correlation with histopathological and microbiological results. *Journal of Clinical Microbiology* **32**, 46–50.

Roseff, S. & Campos, J. (1993). Detection of cytomegalovirus antibodies in serum using the TranSTAT-CMV and CMV scan assays. *American Journal of Clinical Pathology* **99**, 539–541.

Schachter, J. (1985). Immunodiagnosis of sexually transmitted disease. *The Yale Journal of Biology and Medicine* **58**, 443–452.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Schue, V., Green, G., & Monteil, H. (1994). Comparison of the ToxA test with cytotoxity assay and culture for the detection of *Clostridium difficile*-associated diarrhoea disease. *Journal of Medical Microbiology* **41**, 316–318.

Sellors, J., Mahony, J., Jang, D., Pickard, L., Goldsmith, C., Gafni, A., & Chernesky, M. (1991). Comparison of cervical, urethral, and urine specimens for the detection of *Chlamydia trachomatis* in women. *Journal of Infectious Diseases* **164**, 205–208.

Smith, J., Buxton, D., Cahill, P., Fiandaca, M., Goldston, L., Marselle, L., Rigby, S., Olive, D., Hendricks, A., Shimei, T., Klinger, J., Lane, D., & Mahan, D. (1997). Detection of *Mycobacterium tuberculosis* directly from sputum by using a prototype automated Q-beta replicase assay. *Journal of Clinical Microbiology* **35**, 1477–1483.

Staquet, M., Rozencweig, M., Lee, Y., & Muggia, F. (1981). Methodology for the assessment of new dichotomous diagnostic tests. *Journal of Chronic Diseases* **34**, 599–610.

Torrance-Rynard, V. & Walter, S. (1997). Effects of dependent errors in the assessment of diagnostic test performance. *Statistics in Medicine* **16**, 2157–2175.

Walter, S. & Irwig, L. (1988). Estimation of test error rates, disease prevalence and relative risk from misclassified data: a review. *Journal of Clinical Epidemiology* **41**, 923–937.

Wu, C., Lee, M., Yin, S., Yang, D., & Cheng, S. (1991). Comparison of polymerase chain reaction, monoclonal antibody based enzyme immunoassay, and cell culture for detection of *Chlamydia trachomatis* in genital specimens. *Sexually Transmitted Diseases* **19**, 193–197.

Young, H., Anderson, J., Moyes, A., & McMillan, A. (1997). Non-cultural detection of rectal and pharyngeal gonorrhoeae by the Gen-Probe PACE 2 assay. *Genitourinary Medicine* **73**, 59–62.

Zweygberg, W., Landqvist, M., Hokeberg, I., Eriksson, B., Olding-Stenkvist, E., & Grillner, L. (1990). Early detection of cytomegalovirus in cell culture by a new monoclonal antibody. *Journal of Virology Methods* **27**, 211–219.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

APPENDIX

Variances for Estimates using the CRS with Modified Sampling

Large sample variances and covariances for the probabilities of each combination of test results, e.g. p_{+-+}, were obtained from the inverse of the expected Fisher information matrix. The delta method was then used to obtain variance estimates for sensitivity and specificity. Expressions for these variance estimates are given below.

Let $\alpha_{+-} = m_{+-}/n_{+-}$ and $\alpha_{--} = m_{--}/n_{--}$ denote the sampling fractions; i.e. α_{+-} is the fraction of new test positive and culture negative specimens that are tested with the resolver. A sampling fraction equal to one is equivalent to all specimens being tested with the resolver. As expected, both variance expressions given below reduce to the usual binomial variance when both sampling fractions equal one.

The asymptotic variance of the sensitivity estimate is:

$$\begin{aligned}\frac{\gamma}{n\delta^3}\Bigg\{&-\frac{2p_{--+}p_{---}(1+\alpha_{--}p_{--+}-\alpha_{--}+\alpha_{--}p_{---})}{\alpha_{--}(p_{--+}+p_{---})} - 2p_{-+}p_{+--} - 2p_{-+}p_{---} - 2p_{+--}p_{--+}\\ &+\frac{\delta}{\gamma}\left[-\frac{p_{--+}(-\alpha_{--}p_{--+}-p_{---}+\alpha_{--}p_{--+}p_{---}+\alpha_{--}p^2_{--+})}{\alpha_{--}(p_{--+}+p_{---})} - p_{-+}(p_{-+}-1) - 2p_{-+}p_{--+}\right]\\ &+\frac{\gamma}{\delta}\left[-\frac{p_{---}(-p_{--+}+\alpha_{--}p_{--+}p_{---}+\alpha_{--}p^2_{---}-\alpha_{--}p_{---})}{\alpha_{--}(p_{--+}+p_{---})} - 2p_{+--}p_{---}\right.\\ &\left.\left.-\frac{p_{+--}(-p_{+-+}-\alpha_{+-}p_{+--}+\alpha_{+-}p_{+-+}p_{+--}+\alpha_{+-}p^2_{+--})}{\alpha_{+-}(p_{+-+}+p_{+--})}\right]\right\}\end{aligned}$$

where $\gamma = p_{-+} + p_{--+}$ and $\delta = 1 - p_{+--} - p_{---}$. The probability of each combination of test results can be estimated using the expressions given in Section 5.1.

The asymptotic variance of the specificity estimate is:

$$\begin{aligned}\frac{p_{+--}p_{---}}{n(p_{+--}+p_{---})^4}\Bigg\{&\frac{p_{---}(p_{+-+}+\alpha_{+-}p_{+--}-\alpha_{+-}p_{+-+}p_{+--}-\alpha_{+-}p^2_{+--})}{\alpha_{+-}(p_{+-+}+p_{+--})} + 2p_{+--}p_{---}\\ &+\frac{p_{+--}(p_{--+}+\alpha_{--}p_{---}-\alpha_{--}p_{--+}p_{---}-\alpha_{--}p^2_{---})}{\alpha_{--}(p_{--+}+p_{---})}\Bigg\}\end{aligned}$$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Table 1. Contingency tables for summarizing the two stages of discrepant resolution. The number of specimens with each combination of test results is denoted n_{ijk} where i, j, k are the results (+ or −) of the new test, culture, and the resolver, respectively.

		Stage 1		Stage 2	
		Culture		Discrepant Resolution	
		+	−	+	−
New Test	+	n_{++}	n_{+-}	$n_{++} + n_{+-+}$	n_{+--}
	−	n_{-+}	n_{--}	n_{-++}	$n_{--} + n_{-+-}$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

Table 2. Contingency tables summarizing discrepant resolution for the Chlamydia data. EIA is the new test under evaluation, the reference test is cell culture, and the resolver is PCR.

		Stage 1		Stage 2	
		Culture		Resolved using PCR	
		+	−	+	−
EIA	+	20	7	20 + 4	7 - 4
	−	3	294	3 - 1	294 + 1

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/uwbiostat/paper156

Table 3. Contingency tables summarizing the two stages of the composite reference standard. Only culture (C) negative specimens are tested with the resolver (R) in stage 2.

		Stage 1		Stage 2		
		Culture		Composite Reference		
				+		−
		+	−	$C+$	$C-, R+$	$C-, R-$
New Test	+	n_{++}	n_{+-}	n_{++}	n_{+-+}	n_{+--}
	−	n_{-+}	n_{--}	n_{-+}	n_{--+}	n_{---}

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

Table 4. Contingency tables summarizing the composite reference for the Chlamydia data.

		Stage 1		Stage 2	
		Culture		Composite Reference	
		+	−	+	−
EIA	+	20	7	20 + 4	7 - 4
	−	3	294	3 + 2	294 - 2

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/uwbiostat/paper156

Table 5. Comparison of accuracy parameters obtained when three different reference standards are used. The three standards are discrepant resolution (DR), culture, and composite reference (CRS).

Comparison	Prevalence	Sensitivity	Specificity	PPV	NPV
DR vs. Culture	⋆	$\geq$	$\geq$	$\geq$	$\geq$
CRS vs. Culture	$\geq$	⋆	⋆	$\geq$	$\leq$
CRS vs. DR	$\geq$	$\leq$	$\leq$	$=$	$\leq$

⋆ No general statements possible

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

Table 6. Accuracy of EIA using three standards and Chlamydia data. Standard errors are included in parentheses.

Standard	Prevalence	Sensitivity	Specificity	PPV	NPV
Culture	0.071	0.870	0.977	0.741	0.990
	(0.014)	(0.070)	(0.009)	(0.084)	(0.006)
Discrepant Resolution	0.080	0.923	0.990	0.889	0.993
	(0.015)	(0.053)	(0.006)	(0.060)	(0.005)
Composite Reference	0.090	0.828	0.990	0.889	0.983
	(0.016)	(0.070)	(0.006)	(0.060)	(0.007)

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Table 7. Latent class analysis of Chlamydia data. Standard errors are included in parentheses.

Test	Sensitivity	Specificity
EIA	0.909 (0.061)	0.990 (0.006)
Culture	0.834 (0.076)	0.997 (0.003)
PCR	1.000 (0.005)	0.995 (0.005)

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

(a) Estimated sensitivity of EIA.

(b) Estimated specificity of EIA.

Figure 1: Effect of varying n_{--+} on the estimated (a) sensitivity and (b) specificity of EIA when a CRS of culture and PCR is used. The observed estimates are denoted by an asterisk.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

(a) ARE of sensitivity estimate.

(b) ARE of specificity estimate.

Figure 2: Asymptotic relative efficiency (ARE) of (a) sensitivity and (b) specificity when testing only a fraction of the EIA negative/culture negative specimens with PCR to testing all these specimens with PCR.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

(a) ARE of sensitivity estimate.

(b) ARE of specificity estimate.

Figure 3: Asymptotic relative efficiency (ARE) of (a) sensitivity and (b) specificity when testing only a fraction of the EIA positive/culture negative specimens with PCR to testing all these specimens with PCR.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive