STATISTICS IN MEDICINE

USING A COMBINATION OF REFERENCE TESTS
to assess the accuracy of a new diagnostic test

TODD A. ALONZO1* AND MARGARET S. PEPE2

1 Department of Biostatistics, University of Washington, Box 357232, 1705 NE Pacific St., Seattle, WA 98195-7232, U.S.A.
2 Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, MP-665, P.O. Box 19024, Seattle, WA 98109-1024 U.S.A.

SUMMARY
Often the accuracy of a new diagnostic test must be assessed when a perfect gold standard does not exist. Use of an imperfect reference test biases accuracy estimates of the new test. This paper reviews existing approaches to this problem including discrepant resolution and latent class analysis. Deficiencies with these approaches are identified. A new approach is proposed that combines the results of several imperfect reference tests to define a better reference standard. We call this the composite reference standard (CRS). Using the CRS, accuracy can be assessed using multi-stage sampling designs. Maximum likelihood estimates of accuracy and expressions for the variance of sensitivity and specificity are provided. Data from clinical literature on the detection of Chlamydia trachomatis are used to illustrate and compare the different approaches. Advantages of the CRS relative to other approaches include that the CRS is explicitly defined, does not depend on the results of the new test under investigation, and is easy to interpret. Copyright © 1999 John Wiley & Sons, Ltd.

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 an imperfect gold standard. It is generally accepted that culture is nearly 100 per cent specific, but less than 100 per cent sensitive.\(^1,2\) The fact that cell culture is not perfectly specific identifies those with infections as uninfected. Furthermore, this misclassification biases estimates of the accuracy of a new test. If the sensitivity and specificity of culture with respect to a gold standard were known or could be estimated, one could use data comparing the new test to culture in order to indirectly estimate the sensitivity and specificity of the new test with respect to the gold standard.\(^3\) 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 among those 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 standard for evaluating new diagnostic tests, it may be possible to use these more sensitive tests in conjunction with culture as a standard.

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 increasingly 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 based on a single test, we propose the use of a composite reference standard, which is an imperfect gold standard based on multiple tests. 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 gold standard based on a single test (culture), with discrepant resolution, and with the composite reference standard. Some cost efficient study designs for evaluating accuracy using the composite reference standard are considered in Section 5. 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 method 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 a perfect gold standard does not exist, but an imperfect reference test is available. Specifically, DR uses additional ‘resolver’ tests to resolve the discrepant results between the new diagnostic test and the imperfect reference test. For the purpose of discussion
Table I. 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.

<table>
<thead>
<tr>
<th></th>
<th>Stage 1</th>
<th>Stage 2</th>
<th>Discrepant resolution</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Culture</td>
<td></td>
<td></td>
</tr>
<tr>
<td>New test +</td>
<td>$n_{++}$</td>
<td>$n_{++}$</td>
<td>$n_{++}$ + $n_{+-}$</td>
</tr>
<tr>
<td></td>
<td>$n_{-+}$</td>
<td>$n_{-+}$</td>
<td>$n_{-+}$</td>
</tr>
<tr>
<td>New test −</td>
<td>$n_{-+}$</td>
<td>$n_{-+}$</td>
<td>$n_{-+}$ + $n_{--}$</td>
</tr>
<tr>
<td></td>
<td>$n_{--}$</td>
<td>$n_{--}$</td>
<td>$n_{--}$</td>
</tr>
</tbody>
</table>

There are two stages to the discrepant resolution algorithm. In the first stage all $n$ specimens are tested with the new test and with culture. The results of the testing in stage 1 can be summarized in a contingency table (Table I). 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, $n_{+-}$ gives 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 I. 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:

1. \[
\text{prevalence}_C = \frac{n_{++} + n_{+-}}{n}
\]
2. \[
\text{sensitivity}_C = \frac{n_{++}}{n_{++} + n_{+-}}
\]
3. \[
\text{specificity}_C = \frac{n_{-+}}{n_{-+} + n_{--}}
\]
4. \[
\text{PPV}_C = \frac{n_{++}}{n_{++} + n_{+-}}
\]
5. \[
\text{NPV}_C = \frac{n_{-+}}{n_{-+} + n_{--}}
\]

Since culture is known to be an imperfect reference test, DR attempts to improve the standard by re-testing those specimens for which the two tests in the first stage disagree. A resolver test is applied to these $n_{+-}$ and $n_{-+}$ specimens. 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 I). 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 $n_{+-+}$. 
Table II. 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

<table>
<thead>
<tr>
<th></th>
<th>Stage 1</th>
<th></th>
<th>Stage 2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Culture</td>
<td>Resolved using PCR</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>+</td>
<td>−</td>
<td>+</td>
<td>−</td>
</tr>
<tr>
<td>EIA +</td>
<td>20</td>
<td>7</td>
<td>20 + 4</td>
<td>7 − 4</td>
</tr>
<tr>
<td>EIA −</td>
<td>3</td>
<td>294</td>
<td>3 − 1</td>
<td>294 + 1</td>
</tr>
</tbody>
</table>

When the discrepant resolution algorithm is used, the following estimates of prevalence, sensitivity, specificity, PPV, and NPV are obtained:

\[
\text{prevalence}_{\text{DR}} = \frac{n_{++} + n_{+−} + n_{−+}}{n}
\]  
(6)

\[
\text{sensitivity}_{\text{DR}} = \frac{n_{++}}{n_{++} + n_{+−} + n_{−+}}
\]  
(7)

\[
\text{specificity}_{\text{DR}} = \frac{n_{−−}}{n_{−−} + n_{−+} + n_{+−}}
\]  
(8)

\[
\text{PPV}_{\text{DR}} = \frac{n_{++}}{n_{++} + n_{+−} + n_{+−}}
\]  
(9)

\[
\text{NPV}_{\text{DR}} = \frac{n_{−−}}{n_{−−} + n_{−+} + n_{+−}}.
\]  
(10)

2.2. Example

Wu et al.\textsuperscript{28} conducted a study in which specimens from 324 men and women attending two STD clinics in China and Taiwan were tested for \textit{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 II. If culture is considered the standard, then using equation (1) the prevalence of \textit{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 seven specimens which were positive by EIA and negative by culture and the three culture positive/EIA negative specimens are re-tested with the resolver, PCR. Four of the seven EIA positive/culture negative specimens were considered positive by PCR and one of the three EIA negative/culture positives was PCR negative. The results after resolving the discrepancies are summarized in stage 2 of Table II. 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 as \(24/27 = 0.889\) 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. 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 the 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 associated with DR have been described. Miller, Hadgu and Lipman and Astles 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 are biased upwards, so that the new test appears to be more accurate than it really is. Green et al. showed that when the resolver test is not a perfect gold standard even larger biases are possible, but that 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 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 an imperfect gold 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. 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 DR approach while avoiding its most problematic disadvantages. Like DR, the CRS approach allows one: (i) to use several sources of information in order to assess if an infection is present; (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 DR, however, the CRS 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 suggested by Miller and has been used by Jang et al. for the chlamydia setting. To assess the performance of the new test using CRS, there are two stages to the testing (Table III). As with DR, in the first stage all specimens are tested by the new test.
Table III. Contingency table summarizing the two stages of the composite reference standard. Only culture (C) negative specimens are tested with the resolver (R) in stage 2.

<table>
<thead>
<tr>
<th>Stage 1</th>
<th>Stage 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Composite reference</td>
</tr>
<tr>
<td></td>
<td>+</td>
</tr>
<tr>
<td>Culture</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>−</td>
</tr>
</tbody>
</table>

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 III. For example, \( n_{+−} \) is the number of new test positive/culture negative/resolver positive specimens. The results from both stages of testing are combined to get valid estimates of the accuracy of the new test using CRS.

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

\[
sensitivity_{CRS} = \frac{n_{++} + n_{+−}}{n_{++} + n_{+−} + n_{−+} + n_{−−}}
\]  
(11)

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

\[
specificity_{CRS} = \frac{n_{−−}}{n_{−−} + n_{+−}}.
\]  
(12)

Similarly

\[
prevalence_{CRS} = \frac{n_{++} + n_{+−} + n_{−+} + n_{−−}}{n}
\]  
(13)

\[
PPV_{CRS} = \frac{n_{++} + n_{+−}}{n_{++} + n_{+−} + n_{−−}}
\]  
(14)

\[
NPV_{CRS} = \frac{n_{−−}}{n_{−−} + n_{−+} + n_{−−}}.
\]  
(15)

3.2. Example

Again consider the data from the study done by Wu et al.²⁸ A contingency table summarizing the two stages of CRS for these data is given in Table IV. 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 of defining the CRS 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, four of the seven EIA positive/culture negatives were PCR positive. Therefore, an additional six 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.
Table IV. Contingency table summarizing the composite reference for the chlamydia data

<table>
<thead>
<tr>
<th></th>
<th>Stage 1</th>
<th></th>
<th>Stage 2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>EIA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+</td>
<td>20</td>
<td>7</td>
<td>20</td>
<td>4</td>
</tr>
<tr>
<td>−</td>
<td>3</td>
<td>294</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>

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.

### 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 in a real study, we determined (Figure 1) the effect of varying values of $n_{-+}$ on the CRS estimates of sensitivity and specificity in the context of the data from Wu et al.\textsuperscript{28}

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, nine (3.06 per cent) tested positive with the resolver, the estimate of sensitivity drops substantially to 0.667. Thus, the sensitivity estimate depends heavily on the results of the resolver testing of the new test negative/culture negative specimens. On the other hand, Figure 1(b) suggests the specificity estimate does not depend heavily on the testing of these specimens. Estimated specificity decreases slowly with $n_{-+}$. If none of the new test negative/culture negative specimens tests positive with PCR, the specificity is 0.990, while the specificity only drops to 0.980 when 147 (50 per cent) of these specimens test positive with PCR. In practice such a high percentage of positives would be extremely unlikely.
Table V. Comparison of accuracy parameters obtained when discrepant resolution (DR), culture and composite reference standard (CRS) are used

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Prevalence</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>PPV</th>
<th>NPV</th>
</tr>
</thead>
<tbody>
<tr>
<td>DR versus culture</td>
<td>*</td>
<td>≥</td>
<td>≥</td>
<td>≥</td>
<td>≥</td>
</tr>
<tr>
<td>CRS versus culture</td>
<td>≥</td>
<td>*</td>
<td>*</td>
<td>≥</td>
<td>≤</td>
</tr>
<tr>
<td>CRS versus DR</td>
<td>≥</td>
<td>≤</td>
<td>≤</td>
<td>=</td>
<td>=</td>
</tr>
</tbody>
</table>

*No general statements possible

4. COMPARISON OF ACCURACY PARAMETERS

Thus far we have described three standards for assessing the performance of a new diagnostic test: an imperfect gold standard based on a single test (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 the comparisons are summarized in Table V.

Comparing equations (1) and (6) suggests that prevalence of infection when the DR algorithm is used can be lower or higher than prevalence estimated using culture as the standard with equality when \( n_{+} - = n_{+} - + \). On the other hand, sensitivity, specificity, PPV and NPV of the new test when DR is used is always greater 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 when the CRS is used, prevalence of infection is always the same or greater than prevalence when culture alone is used as the standard. Furthermore, since prevalence estimated using the CRS is always greater than or equal to the culture based estimate of prevalence, it makes sense that PPV estimated using the CRS is greater than or equal to the culture based estimate of PPV, and NPV estimated using the CRS is less than or equal to the 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, that is, 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:

\[
A: \text{sensitivity}_C \geq \text{sensitivity}_{\text{CRS}}, \quad \text{if and only if} \quad \text{sensitivity}_C \geq \frac{n_{+} - +}{n_{+} - + + n_{-} - +}.
\]

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 \( n_{-} - +/(n_{+} - + + n_{-} - +) \):

\[
B: \text{specificity}_C \leq \text{specificity}_{\text{CRS}}, \quad \text{if and only if} \quad \text{specificity}_C \geq \frac{n_{-} - +}{n_{+} - + + n_{-} - +}.
\]
Table VI. Accuracy of EIA using three standards and the chlamydia data. Standard errors are included in parentheses

<table>
<thead>
<tr>
<th>Standard</th>
<th>Prevalence</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>PPV</th>
<th>NPV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Culture</td>
<td>0.071</td>
<td>0.870</td>
<td>0.977</td>
<td>0.741</td>
<td>0.990</td>
</tr>
<tr>
<td></td>
<td>(0.014)</td>
<td>(0.070)</td>
<td>(0.009)</td>
<td>(0.084)</td>
<td>(0.006)</td>
</tr>
<tr>
<td>Discrepant resolution</td>
<td>0.080</td>
<td>0.923</td>
<td>0.990</td>
<td>0.889</td>
<td>0.993</td>
</tr>
<tr>
<td></td>
<td>(0.015)</td>
<td>(0.053)</td>
<td>(0.006)</td>
<td>(0.060)</td>
<td>(0.005)</td>
</tr>
<tr>
<td>Composite reference</td>
<td>0.090</td>
<td>0.828</td>
<td>0.990</td>
<td>0.889</td>
<td>0.983</td>
</tr>
<tr>
<td></td>
<td>(0.016)</td>
<td>(0.070)</td>
<td>(0.006)</td>
<td>(0.060)</td>
<td>(0.007)</td>
</tr>
</tbody>
</table>

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. 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 PPV estimates resulting from the two methods are the same.

In Table VI, using data from Wu et al., estimated accuracy and prevalence parameters calculated using culture, DR and CRS are displayed. The inequalities in Table V are borne out with the real data. The estimate of sensitivity using CRS is less than the culture based estimate, whereas specificity is greater when the CRS is used. Of the six specimens testing negative with culture and positive with the resolver only four test positive with the new test. Thus the right hand side of statements A and B held, implying the left hand side inequalities.

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 needs 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 the accuracy parameters associated with the new test using the CRS can be calculated based on data from such a study design.

5.1. Modified 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 likelihood for these data is:

\[ c_0 p_{++}^{n_{++}} p_{+-}^{n_{+-}} p_{-+}^{m_{+-}} p_{--}^{m_{--}} \]

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:

\[
\hat{p}_{++} = \frac{n_{++}}{n} \quad \hat{p}_{-+} = \frac{n_{-+}}{n} \\
\hat{p}_{+-} = \frac{m_{+-} (n_{+-}/m_{+-})}{n} \quad \hat{p}_{--} = \frac{m_{--} (n_{--}/m_{--})}{n} \\
\hat{p}_{-+} = \frac{m_{-+} (n_{-+}/m_{-+})}{n} \quad \hat{p}_{--} = \frac{m_{--} (n_{--}/m_{--})}{n}.
\]

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 (that is, the numerators of \( \hat{p}_{++}, \hat{p}_{-+}, \hat{p}_{+-} \) and \( \hat{p}_{--} \)) can be interpreted as the observed cell frequencies multiplied by the inverse of the sampling fraction. Inserting these estimates into expressions for prevalence and accuracy yields the following:

\[
\text{prevalence}_{\text{CRS}} = \frac{n_{++} + m_{+-} (n_{+-}/m_{+-}) + n_{-+} + m_{--} (n_{--}/m_{--})}{n} \\
\text{sensitivity}_{\text{CRS}} = \frac{n_{++} + m_{+-} (n_{+-}/m_{+-})}{n_{++} + m_{+-} (n_{+-}/m_{+-}) + n_{-+} + m_{--} (n_{--}/m_{--})} \\
\text{specificity}_{\text{CRS}} = \frac{m_{--} (n_{--}/m_{--})}{m_{--} (n_{--}/m_{--}) + m_{+-} (n_{+-}/m_{+-})} \\
\text{PPV}_{\text{CRS}} = \frac{n_{++} + m_{+-} (n_{+-}/m_{+-})}{n_{++} + m_{+-} (n_{+-}/m_{+-}) + m_{+-} (n_{+-}/m_{+-})} \\
\text{NPV}_{\text{CRS}} = \frac{m_{--} (n_{--}/m_{--})}{m_{--} (n_{--}/m_{--}) + n_{++} + m_{+-} (n_{+-}/m_{+-})}.
\]

This method of reweighting using the sampling fractions is similar to that of Irwig et al.\(^{35}\)

### 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 calculating 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\(^{36}\) compares different methods for calculating confidence intervals for a proportion including using a logit transformation.
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

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 sensitivity and specificity 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 the resolver, that is, the asymptotic relative efficiency (ARE). In these plots the sampling fraction, \( m_{+--}/n_{+--} \), is set to 1, that is, all specimens testing positive with EIA and negative with the culture are tested with the PCR resolver. It appears that the sampling fraction \( m_{--}/n_{--} \) has little effect on the precision with which 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. With Wu’s data, when 50 per cent of the culture negative/EIA negative specimens are re-tested the standard error of the sensitivity estimate is 15 per cent larger than when all \( n_{--} \) specimens are re-tested with the resolver.

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

6. LATENT CLASS ANALYSIS

A statistical solution to the imperfect gold standard problem put forth by Walter and Irwig, among 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 and the CRS approach. Maximum likelihood
techniques then yield estimates of prevalence of infection 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. 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. Finally, the estimates calculated using this approach are not simple explicit functions of the data. Thus for the clinician, the estimates are output from a 'black box' statistical algorithm. Alonzo and Pepe\(^{38}\) provide analytic expressions for estimates of prevalence, sensitivity and specificity resulting from latent class analysis which help give an intuitive understanding for how the data affect the estimates. These expressions allow one to investigate how the violation of the conditional independence assumption can affect parameter estimation. Alonzo and Pepe\(^{38}\) show 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 and Walter.\(^{39}\)

The latent class method was applied to the data from Wu \textit{et al.}\(^{28}\) This method estimated prevalence of chlamydia to be 0.081 and yielded the estimates of sensitivity and specificity shown in Table VII for each of the three tests, EIA, culture and PCR. Comparing these results to those given in Table VI suggests that the prevalence estimate given by latent class analysis is in between the estimates obtained when culture is considered the standard and when the CRS is used. The

Table VII. Latent class analysis of the chlamydia data. Standard errors are in parentheses

<table>
<thead>
<tr>
<th>Test</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>EIA</td>
<td>0.909 (0.061)</td>
<td>0.990 (0.006)</td>
</tr>
<tr>
<td>Culture</td>
<td>0.834 (0.076)</td>
<td>0.997 (0.003)</td>
</tr>
<tr>
<td>PCR</td>
<td>1.000 (0.005)</td>
<td>0.995 (0.005)</td>
</tr>
</tbody>
</table>
estimate of sensitivity for EIA resulting from latent class analysis is higher than that obtained using culture or the CRS 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 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 accuracy of a new diagnostic test when a gold standard does not exist. Deficiencies associated with the use of a single imperfect reference test, with discrepant resolution, and with 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 it ascertains 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 standard 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 gold standard. The CRS is likely more palatable than culture alone, and thus is a viable alternative to DR. As it turns out, the idea of a CRS has been suggested in the statistical literature and several different composite reference standards have already been used to study the performance of new tests to detect Chlamydia trachomatis along with other infections such as Bordetella pertussis. Some researchers have used a CRS defined just as we have. 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 resolver tests are considered CRS positive. A contribution of this paper is to formalize the notion of a CRS and to suggest how accuracy can be estimated statistically when a CRS is used with 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 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 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. have shown that DR estimates specificity reasonably well but does not estimate sensitivity well. Our results also show that inference about sensitivity but not specificity depends on the 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 specimens from re-testing.

The modified sampling design, for use with the CRS, discussed in Section 5 is similar to a double sampling design in that a random sample of specimens is selected for additional testing. In a double sampling design, all specimens receive an imperfect test, and some specimens also receive a perfect test. However, in the chlamydia setting a perfect test does not exist. Therefore, in our modified sampling design, all specimens receive one imperfect test (culture), and some specimens receive another imperfect test (resolver test). Results from both imperfect tests are then used to define the CRS.

Although the estimates obtained using the DR 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 interpreting the results.

As technological 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 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.

APPENDIX: VARIANCES FOR ESTIMATES USING THE CRS WITH MODIFIED SAMPLING

Large sample variances and covariances for the probabilities of each combination of test results, for example, $p_{+--}$, 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 $x_{+} = m_{+}/n_{++}$ and $x_{--} = m_{--}/n_{--}$ denote the sampling fractions, that is, $x_{+}$ is the fraction of new test positive/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

$$\frac{\gamma}{n\delta^2} \left\{ -2p_{--}^2(1 + x_{--}p_{--} - x_{--} + x_{--}p_{--} - 2p_{--}p_{++}) \right\}$$

$$-2p_{+}p_{--} - 2p_{++}p_{--}$$
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

\[
\frac{p_{++} - p_{--}}{n(p_{++} + p_{--})^2} \left\{ \frac{p_{--}(p_{++} + p_{--} - p_{++} - p_{--} - p_{++}^2)}{\gamma(p_{++} + p_{--})} + 2p_{++}p_{--} + \frac{p_{++}(p_{++} + p_{--} - p_{++} - p_{--} - p_{++}^2)}{\gamma(p_{++} + p_{--})} \right\}.
\]

ACKNOWLEDGEMENTS

We thank Linda Weisman for help in typing the manuscript. Support for this research was provided by NIH grant R01 GM54438.

REFERENCES


