# Some Issues in Resolution of Diagnostic Tests using an Imperfect Gold Standard

**Technical Report 628**
**School of Statistics**
**University of Minnesota**

Published in; *Statistics in Medicine*, 20, (2001), 1987—2001.

Douglas M. Hawkins
School of Statistics
University of Minnesota
Minneapolis,
MN 55455-0493
Tel (612) 624-4166
e-mail doug@stat.umn.edu

James A. Garrett
Becton Dickinson Bioscience
Sparks
MD 21152-0999
e-mail Jim_Garrett@ms.bd.com

Betty Stephenson
Becton Dickinson Bioscience
Sparks
MD 21152-0999
e-mail Betty_Stephenson@ms.bd.com

# Some Issues in Resolution using an Imperfect Gold Standard


## Summary

As a subject's true disease status is seldom known with certainty, it is necessary to compare the performance of new diagnostic tests with those of a currently accepted but imperfect "gold standard." Errors made by the gold standard mean that the sensitivity and specificity calculated for the new test are biased, and do not correctly estimate the new method's sensitivity and specificity. The traditional approach to this problem was "discrepant resolution," in which the subjects for whom the two methods disagreed were subjected to a third "resolver" test. Recent work has pointed out that this does not automatically solve the problem. A sounder approach goes beyond the discordant test results and tests at least some of the subjects with concordant results with the resolver also. This leaves some issues unresolved. One is the basic question of the direction of biases in various estimators. We point out that this question does not have a simple universal answer. Another issue, if one is to test a sample of the subjects with concordant results rather than all cases, is how to compute estimates and standard errors of the measures of test performance, notably sensitivity and specificity of the test method relative to the resolver. Expressions for these standard errors are given and illustrated with a numeric example. It is shown that using just a sample of subjects with concordant results may lead to great savings in assays. The design issue of how many concordant cells to test depends on the numbers of concordants and discordants. The formulas given show how to evaluate impact of different choices for these numbers and hence settle on a design that gives the required precision of estimates.


**Keywords:** Bias, diagnostic test, misclassification, sensitivity, specificity, diagnostic performance, test evaluation.

**Introduction**

A common problem faced in medical diagnostic testing is that of determining the diagnostic performance of a qualitative "index test," that is, a test being evaluated that yields a positive or negative result rather than a quantitative measurement. Conceptually (though perhaps only with great difficulty operationally), subjects can be divided into those that have do indeed have the condition being tested for, and those that do not. We will refer to the former group as 'disease positive' and the latter as 'disease negative.' The performance of a test is usually summarized by its sensitivity and specificity, sensitivity being the proportion of disease positive subjects that yield positive test results, and specificity being the proportion of disease negative subjects that yield negative test results.

In an ideal world, one can estimate sensitivity and specificity by applying the index test to a series of subjects whose true disease status can be ascertained. Estimation of sensitivity and specificity is straightforward in this case; and likewise the associated confidence intervals are easy to calculate (see for example Fleiss 1981). It is more often the case, however, that perfect categorization of the subjects is impossible, unethical, or too expensive, and one must settle for comparison with an imperfect "reference" method. Misclassification by the reference method introduces biases into the sensitivity and specificity estimates. These biases are usually downward and under some circumstances can be severe: Valenstein (1990) offers an example in which an index test's true sensitivity of 98% would appear to be 67.1%, a downward bias of 30.9 percentage points.

One attempt to address these biases has been through "discrepant analysis," or "discrepant resolution," in which subjects yielding discordant test results by the index and reference method are tested by a third "resolving" method which may itself be either imperfect or perfect. Variations on this design have been discussed such as applying the resolver test only to apparently false negatives (those positive by the index test but negative by the reference test) or only to apparently false positives (those with the opposite discordancy). While discrepant analysis of any form is intended to yield additional information about potentially problematic subjects, it introduces its own set of biases, always in an upward direction (Hadgu, 1996). There has been some debate on the questions of whether these biases more or less counteract misclassification biases, and whether discrepant resolution is the lesser of two evils (see Hadgu, 1997; and Green et al., 1998). In this paper, we further explore the origins of the biases introduced by misclassification and classical discrepant analysis, and find that they do not permit simple summaries.

A number of authors have proposed model-based estimates, or estimates that make use of prior information, to avoid misclassification bias without retesting any subjects. The literature on one such method, Latent Class analysis (LCA), is large. Those who have applied LCA to diagnostic tests include Hui and Walter (1980), Joseph et al. (1995), and Rindskopf and Rindskopf (1986). Others, such as Staquet et al. (1981), have developed mathematical adjustments to estimates that can be applied if certain quantities are regarded as "known," such as the sensitivity of the reference test. All of these efforts assume that index and reference tests err independently (that is, are conditionally independent given true disease status). This is a highly questionable assumption in most cases. Recently Qu et al. (1996), Qu and Hadgu (1998), Hadgu and Qu (1998), and Shih and Albert (1999) have generalized LCA in an effort to allow for the more realistic assumption of conditional dependence between the two methods being compared.

Meier (1998) suggested a modified form of discrepant analysis whereby the resolving method is applied to a random sample of subjects with concordant results in addition to those with discrepant results. This idea is appealing, but no method of statistical analysis for this design appears to have been documented. We address this by proposing a method for estimating sensitivity and specificity. The binomial distribution of positive test results then leads to standard errors and approximate confidence intervals for the test performance measures. This method does not assume conditional independence, nor is it based on a latent class model. We apply the method to one constructed and one real example, and show that this study design, analyzed with our proposed method, makes efficient use of resolving tests.

**Estimating Sensitivity and Specificity When Subject Status Is Known**

Suppose an index test is applied to a group of subjects whose true disease status is known. Setting up the 2 × 2 cross-classification of the subjects' test results by their true disease status leads to the unbiased estimation of the test's sensitivity and specificity. Under a binomial sampling model, this also yields standard errors of the estimates. Confidence intervals can also be calculated in several different ways (see for example Agresti and Coull 1998 for an illuminating discussion.)

In symbols, write the classification table by true disease status

Table 1: Conceptual tabulation of test result by true status (D+ for diseased and D- for not diseased).

| | True status | | |
|---|---|---|---|
| Index test | D+ | D- | Total |
| Positive | $n_{1D}$ | $n_{1N}$ | $n_{1+}$ |
| Negative | $n_{0D}$ | $n_{0N}$ | $n_{0+}$ |
| Total | $n_{+D}$ | $n_{+N}$ | $n_{++}$ |

(following common practice, replacing a subscript by + indicates summing over that subscript so, for example, $n_{1+} = n_{11}+n_{10}$) The estimated sensitivity and its estimated standard error are given by

$$Sens = n_{1D} / n_{+D}, \quad SE = \sqrt{\frac{n_{1D} n_{0D}}{n_{+D}^3}} \tag{1}$$

and the corresponding formulas for specificity are

$$Spec = n_{0N} / n_{+N}, \quad SE = \sqrt{\frac{n_{0N} n_{1N}}{n_{+N}^3}} \tag{2}$$

In reality, it is seldom possible or advisable to rely completely on subjects whose true disease status is known. There are three common reasons for this. One reason is that it is sometimes possible but difficult or expensive to establish a subject's true disease status – as for example in the syphilis

screening example of Irwig et al. (1994). The implication of this is that at most some of the subjects can be diagnosed exactly. The second reason is that there may be no generally accepted and/or definitive way of determining whether a subject does or does not have the disease. A third possible concern is verification bias (Begg and Greenes, 1983).

Absent known true diagnoses, then, one is forced to use an imperfect gold standard as the reference. This contaminates the 'true table,' giving the observed table.

Table 2: Tabulation of subjects by index and reference test

| | Reference test | | |
|---|---|---|---|
| Index test | Positive | Negative | Total |
| Positive | $n_{11}$ | $n_{10}$ | $n_{1+}$ |
| Negative | $n_{01}$ | $n_{00}$ | $n_{0+}$ |
| Total | $n_{+1}$ | $n_{+0}$ | $n_{++}$ |

The 'relative' sensitivity and specificity of the index test – relative, that is, to the imperfect gold standard, are

$$\textit{Relative sensitivity} = n_{11} / n_{+1}$$
$$\textit{Relative specificity} = n_{00} / n_{+0}$$

The contamination of the true table by errors in the reference method results in biased estimates of sensitivity and specificity. These biases however are not at all simple to characterize – not even as to their direction. The difficulty in determining the potential biases can be illustrated by subdividing the subjects by the result of the reference test, which gives Table 3.

Table 3: Conceptual 2x2x2 tabulation of subjects by both tests and true status

| | | Truth | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | D+ | | | D- | | | |
| | | Reference test | | | Reference test | | | |
| | | Positive | Negative | | Positive | Negative | | Total |
| Index Test | Positive | $n_{1D}$ - r | r | $n_{1D}$ | t | $n_{1N}$ - t | $n_{1N}$ | $n_{1+}$ |
| | Negative | $n_{0D}$ - s | s | $n_{0D}$ | u | $n_{0N}$ - u | $n_{0N}$ | $n_{0+}$ |
| | Total | $n_{+D}$ - (r+s) | r+s | | t+u | $n_{+N}$ - (t+u) | $n_{+N}$ | $n_{++}$ |

The letters *r, s, t* and *u* represent errors by the reference method; *r* and *s* are false negatives; *t* and *u* are false positives. Collapsing over the true disease status results in the tabulation shown as Table 4.

Table 4: Observed cross-tabulation by the two tests

| | | Reference test | | Total |
|---|---|---|---|---|
| | | Positive | Negative | |
| Index Test | Positive | $n_{11}=n_{1D}$ - r+t | $n_{10}=n_{1N}$+r - t | $n_{1+}$ |
| | Negative | $n_{01}=n_{0D}$ - s+u | $n_{00}=n_{0N}$+s - u | $n_{0+}$ |
| | Total | $n_{+1}=n_{+D}$ - (r+s)+(t+u) | $n_{+0}=n_{+N}$+(r+s) - (t+u) | $n_{++}$ |

Using the results from Table 4, the relative sensitivity and specificity of the index test relative to the reference test can be written

$$\begin{aligned} \textit{Relative sensitivity} &= (n_{1D} - r+t) / [n_{+D} - r+t - s+u] \\ \textit{Relative specificity} &= (n_{0N}+s - u) / [n_{+N}+r - t+s - u] \end{aligned} \qquad (3)$$

As these expressions illustrate, the connections between the true and the relative sensitivity and specificity are not straightforward. Depending on how errors in the reference method distribute among the four erroneous cells, the estimated sensitivity and specificity may be biased either upward or downward. It is even possible for the net errors *t* - *r* and *s* - *u* to be zero, resulting in unbiased sensitivity and specificity estimates. Sweeping generalizations about the bias in the estimated sensitivity and specificity, therefore, require considerable caution.

**Using a resolver test in 'discrepant analysis'**

The presence of bias in the estimates of sensitivity and specificity leads to the idea of using a further test to locate and correct the errors in the reference method. Conventional 'discrepant analysis' (see for example Hadgu 1997 for discussion and a numeric example) tests the subjects in cells where the index test was positive and the reference test negative and uses this to calculate a 'resolved' sensitivity $(n_{11}+r)/(n_{+1}+r)$. A similar resolution where the index test was negative and the reference test positive is used in the same way to get a 'resolved' specificity $(n_{00}+s)/(n_{+0}+s)$.

As has been pointed out recently (Hadgu 1996, 1997, 1998, Miller 1998), this approach does not remove biases in the estimated sensitivity and specificity. Using a perfectly specific resolver test to resolve the status of the subjects in the top right cell corrects only the *r* cases in which the reference method gave a false negative. It will not identify the *t* cases in which the reference method gave a false positive, or the *u* cases where the index test was falsely negative. Because the bias involves all four quantities *r, s, t* and *u*, knowing the value of only one of these quantities cannot lead to an unbiased estimate of sensitivity.

Similarly, the use of a perfectly sensitive resolver test on the lower left cell of the table does not lead to an unbiased estimate of specificity.

While discrepant analysis leads to 'resolved' sensitivity and specificity estimates that are always higher than the relative sensitivity and specificity, it is not automatically true that it is more biased than the relative figures. In different circumstances either of these potentially biased estimates may be closer to the true sensitivity and specificity and neither is guaranteed to be the less biased.

**Using a perfect resolver to produce consistent sensitivity and specificity estimates**

Conventional discrepant analysis can not be relied upon to produce unbiased estimates of the index method's true sensitivity and specificity. As the expanded table (Table 3) shows, producing unbiased estimates of the index method's sensitivity and/or specificity requires either the exact values or estimates of the differences *(t - r)* and *(s - u)*. These could be found directly from the full 2×2×2 table showing the subject's true status along with the classification by the index and the reference tests, if this were available. Note that this may be thought of as 'resolving' not just the discordant cells of the original 2×2 table, but also the concordant cells.

Write $P_{ijk}$ for the true probabilities in the 2×2×2 table. The index $i$ refers to the index test, $j$ to the reference test, and $k$ to the true status. The value '1' for $i$, $j$ and $k$ indicates a positive test, or a diseased status; and the value '0' indicates a negative test, or a non-diseased status. Use a '+' sign in place of a subscript to indicate a marginal total over that subscript. For example $P_{ij+}$ refers to the cross tabulation by the index and reference tests, without regard to the true disease status.

The prevalence of disease is given by $P_{++1}$, the marginal probability that a subject is diseased. The index test has a true sensitivity of $P_{1+1}/P_{++1}$ and a true specificity of $P_{0+0}/P_{++0}$.

Estimates of the 8 probabilities in the 2×2×2 table can be found by taking a total of $n$ subjects and testing them using the index and reference tests and also the perfect resolver. Write $n_{ijk}$ for the number of subjects in the $i, j, k$ cell of the table. Estimates of $p_{ijk}$ and its standard error are given by

$$p_{ijk} = \frac{n_{ijk}}{n}, \quad SE = \sqrt{\frac{p_{ijk}(1-p_{ijk})}{n}} \tag{4}$$

We get consistent estimates and standard errors of the true sensitivity and specificity of the index test by adding across the levels of the reference test.

**Random sampling approaches**

This scenario is unrealistic, however. If the perfect resolver could reasonably be used on all subjects, it would likely be used as the reference method. Often it is not used as the reference because it is difficult or expensive to apply. In this case, it cannot be used on all *n* subjects but it could reasonably be used on a subset of them. This raises the possibility of randomly sampling from the cells of the 2×2 tabulation of the index test by the reference test, and evaluating those subjects with the perfect resolver.

The observed frequencies of the collapsed 2×2 table defined by the index and reference tests are:

| | | Reference Test | | |
|---|---|---|---|---|
| | | Positive | Negative | Total |
| Index Test | Positive | $n_{11+}$ | $n_{10+}$ | $n_{1++}$ |
| | Negative | $n_{01+}$ | $n_{00+}$ | $n_{0++}$ |
| | Total | $n_{+1+}$ | $n_{+0+}$ | n |

The estimates of the 2×2 marginal probabilities $P_{ij+}$ and standard errors are given by:

$$p_{ij+} = \frac{n_{ij+}}{n}, \quad SE = \sqrt{\frac{p_{ij+}(1-p_{ij+})}{n}} \tag{5}$$

We now test some possibly smaller number $m_{ij}$ of these $n_{ij+}$ subjects using the perfect resolver, and find that a proportion $r_{ij}$ of these are diseased. Then $r_{ij}$ estimates the conditional probability that a subject is diseased given that the subject is classified in cell *i,j* by the index and reference tests. Its estimated standard error is given by

$$SE = \sqrt{\frac{r_{ij}(1-r_{ij})}{m_{ij}}} \tag{6}$$

This estimate of the conditional probability of disease given classification *i,j* and the estimate of the marginal probability $P_{ij+}$ can be multiplied to get an estimate of the joint probability $P_{ij1}$. The standard error of this estimate can be found approximately using the 'delta' method as

$$(SE_{ij1})^2 = (p_{ij+})^2 \frac{r_{ij}(1-r_{ij})}{m_{ij}} + (r_{ij})^2 \frac{p_{ij+}(1-p_{ij+})}{n} \tag{7}$$

From these estimates and standard errors, we can derive those of the sensitivity and specificity. The probability $P_{1+1}$ of the index method giving a true positive result is estimated as

$$p_{11+}r_{11} + p_{10+}r_{10} \tag{8}$$

The probability $P_{0+1}$ of the index method giving a false negative result is estimated as

$$p_{01+}r_{11} + p_{00+}r_{00} \tag{9}$$

The true sensitivity can then be estimated by (see also equation 3.5 of Begg and Greenes 1983)

$$(p_{11+}r_{11} + p_{10+}r_{10})/\,(p_{11+}r_{11} + p_{10+}r_{10} + p_{01+}r_{11} + p_{00+}r_{00}) \tag{10}$$

**Standard errors for derived quantities**

While the estimates of true sensitivity and specificity can be derived in a straightforward manner, derivation of their standard errors is more complicated because it is necessary to take into account the correlation between the different cells in the table. Writing

$$X = p_{11+}r_{11} + p_{10+}r_{10}$$

for the estimated probability of a true positive result by the index method and

$$Y = p_{11+}r_{11} + p_{10+}r_{10+} + p_{01+}r_{11} + p_{00+}r_{00}$$

for the estimated prevalence of the disease, the estimated sensitivity $X/Y$ (expressed as a proportion, not a percentage) has a sampling variance

$$Var\left(\frac{\mathrm{X}}{\mathrm{Y}}\right) = \frac{Var(\mathrm{X})}{[E(\mathrm{Y})]^2} + \frac{[E(\mathrm{X})]^2}{[E(\mathrm{Y})]^4} Var(\mathrm{Y}) - 2\frac{E(\mathrm{X})}{[E(\mathrm{Y})]^3} Cov(X,Y) \tag{11}$$

(details of the calculations are given in the Appendix). An approximate confidence interval can be generated for sensitivity (or specificity) using the usual ('Wald') normal approximation 'estimate $\pm$ two standard errors.' Though as noted in Agresti and Coull (1998) this common confidence interval does not have very well-controlled coverage for small samples, particularly if the sensitivity is close to 1. They recommend the "adjusted Wald" procedure, in which, in each calculation of a proportion or a standard error, two artificial positives and two artificial negatives are added to the sample, and which has much better-controlled coverage.

**Numeric example**

A numeric example may help to clarify the plethora of symbols. The following numbers were constructed to illustrate the methods of calculation. We have 3,000 subjects classified by an index test and an imperfect reference test. Table 5 shows the frequencies in each cell, along with that cell's percentage of the full set of 3,000 subjects.

Table 5. Cross-tabulation of frequencies in the numeric example

| | | Reference test | | |
|---|---|---|---|---|
| | $n_{ij+}$ | Positive | Negative | Total |

| | | | | |
|---|---|---|---|---|
| Index test | Positive | 1800 (60%) | 600 (20%) | 2400 (80%) |
| | Negative | 200 (6.67%) | 400 (13.33%) | 600 (20%) |
| | Total | 2000 (66.67%) | 1000 (33.33%) | 3000 (100%) |

Suppose we decide to test 100 randomly selected specimens from each of the four cells in this table with a perfect resolver, and get the following positives by the resolver:-

Table 6: Results of applying the resolver test to 100 subjects in each cell

| Resolver Test Positive | | | |
|---|---|---|---|
| | | Reference test | |
| | *count* ($r_{ij}$) | Positive | Negative |
| Index test | Positive | 95 | 40 |
| | Negative | 15 | 5 |

Using the proportions computed in Tables 5 and 6, estimates of the joint probabilities ($p_{ij1} = p_{ij+}\, r_{ij}$) can be found. This leads to the estimates of the probabilities of the full 2×2×2 table, given as Table 7.

Table 7: Estimated 2x2x2 table of probabilities

| | | Resolver test ("Truth") | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Positive | | | Negative | | | |
| | | Reference test | | | Reference test | | | |
| | | Positive | Negative | Total | Positive | Negative | Total | Total |
| Index Test | Positive | 0.57 | 0.08 | 0.65 | 0.03 | 0.12 | 0.15 | 0.80 |
| | Negative | 0.01 | 0.0067 | 0.0167 | 0.0567 | 0.1267 | 0.1833 | 0.20 |
| | Total | 0.58 | 0.0867 | 0.6667 | 0.0867 | 0.2467 | 0.3333 | 1.00 |

Using the values in Table 7, we can estimate several useful quantities:

| | |
|---|---|
| Proportion true positive by the index test | 0.57 + 0.08 = 0.65. |
| Proportion false negative by the index test | 0.01 + 0.0067 = 0.0167. |
| Prevalence of disease | 0.65 + 0.0167 = 0.6667 |
| Sensitivity of the index test | 0.65/0.6667 = 0.975 (or 97.5%). |

The first of these is the estimated proportion of the entire population who both have the disease and test positive by the index test. This is the quantity $X$ of equation (11). The second is the proportion of the entire population who have the disease but do not test positive by the index test. This correspond to $Y$-$X$ in equation (11). Note the discrepancy between the estimated sensitivity of the index test as computed

using the perfect resolver information (97.5%) and the sensitivity of the index test relative to the [imperfect] reference test of (100 × 1800/2000) = 90%.

In order to calculate a confidence interval for sensitivity, we need to calculate the standard error based on (11). The intermediate calculations for the standard error are shown in the Appendix.

The calculated value of the standard error of the sensitivity estimate is found to be 0.00567. A 95% confidence interval for sensitivity (using the usual Wald normal approximation) is:

$$\text{Sensitivity} = 0.975 \pm 2 \times 0.00567 = 0.975 \pm 0.011$$

This standard error (and therefore confidence interval width) matches what could be found from a sample of size $0.975(1-0.975)/0.00567^2 = 758$ disease positives assayed by the index test. With a prevalence of 67%, if we had validated the index test directly against the resolver, to attain this precision would have required n=1130 samples analyzed by the index and the resolver tests. The actual number of resolver assays used, 400, is barely one third this number. If the resolver were a substantially more difficult or expensive assay than the reference, this could lead to a major simplification and saving. The source of this saving is the stratification that the original 2×2 table provides so that relatively little investment in the resolver will produce high precision in the final estimates.

The confidence interval above is produced by the 'Wald' formula 'estimate $\pm$ 2 standard errors.' If concerned about the possible inadequacy of this standard approach, we might try the modified Wald approach in which all the estimates and standard errors are computed after adding two artificial positives and negatives to each of the 8 tests. We omit the details of this calculation (they exactly parallel those of the appendix) – they give rise to an estimated sensitivity of 0.9697 with standard error 0.00628 for a nominal 95% confidence interval of (0.957, 0.982). This compares with the traditional interval of (0.964, 0.986). In this example the difference between the two is very slight but this is not the case in situations in which the numbers of positives and negatives are small.

**Use of different sample numbers in the cells**

This sample calculation was based on the assumption of equal numbers of tests sampled in the four cells, leaving open the question of whether it would be better to allocate total resolver sample size unequally between the four cells. This is indeed the case. A standard formulation is (see for example Cochran (1977), page 96-99) to pick the sample allocation that would minimize the standard error of the final estimate of test sensitivity. While we have an explicit formula for this standard error, it depends on not only the $m_{ij}$ but also the proportions $p_{ij}$ and $r_{ij}$, the latter of which are known only after the trial has been completed. We can however use the numbers of this example to illustrate the improvements that could be made by using unequal sample sizes in the four cells. Using the $p_{ij}$ and $r_{ij}$ values and searching for the best allocation of 400 total resolver samples between the $m_{ij}$ gives the values

$$m_{11} = 23,\ m_{10} = 17,\ m_{01} = 162,\ m_{00} = 198$$

which minimize the standard error of the sensitivity. This optimization was done subject to the constraints that the total resolver sample size was 400, and $m_{ij} <= n_{ij}$ for all cells.

The resulting standard error of the estimated sensitivity is 0.00447. This represents an efficiency improvement of $(0.00567/0.00443)^2 = 1.63$, or a 63% improvement over the equal-number allocation. In sample number terms, this means that this allocation of resolver assays would give a sensitivity as precise as a 63% larger equally-split sample. A saving this large is worth some effort to find, and leads to ideas such as a two-stage approach of taking a small random sample from each cell to get preliminary estimates of the $r_{ij}$ so that these can be used for more purposeful allocation of the remaining total sample size.

It should be noted that the unequal allocation gives a much more precise estimate of sensitivity, but this comes at the cost of a much less precise specificity.

This calculation is reminiscent of one done by Irwig et al. (1994) in the more straightforward case of a single test against a gold standard without the additional complication of a third assay.

Our numeric example was made up, and illustrates the situation in which the prevalence was neither very high nor very low, leading to large values in all cells of the contingency table and the opportunity for subsampling in all cells. It is more common for the subject pool to have a disease prevalence close to either zero or 100%, when typically only one cell of the table will have so large a frequency as to invite subsampling. This situation is simpler than the one we illustrated; one will resolve all the subjects in the low-frequency cells and use subsampling on the high frequency cell only. The calculations though will carry over in exactly the same way. The consequences of error in diagnostic settings are different from those in screening, and this determines whether it is the sensitivity or the specificity that should be estimated with more precision.

**Resolution in situations with no perfect resolver**

In many applications, there is no perfect resolver, but there are other imperfect methods that may be used to bring further light to the true status of the specimens.

One attractive method when there is no perfect gold standard is the 'composite reference standard' (CRS) of Alonzo and Pepe (1999). Suppose the reference test is highly specific but not very sensitive, as in Alonzo and Pepe's *Chlamydia trachomatis* example. Then a number of the reference test negative specimens may in fact be true positives. The CRS then defines a subject as positive if either of the tests is positive, and as negative only if both tests are negative. To the extent that the reference and the resolver test give results that are statistically independent conditional on the sample, the CRS can be expected to be both highly specific and quite sensitive. Suppose for example that the reference and resolver test both have specificity 99% and sensitivity 90%, and are conditionally independent. Then the CRS will have a specificity of $100 \times 0.99^2\% = 98\%$ and a sensitivity of $100(1\text{-}0.1^2)\% = 99\%$. Paralleling Alonzo and Pepe, our methods may be used here too to subsample the cells of the original 2×2 table to produce estimates of the probabilities of the original 2×2×2 table of all three tests. For example, if the CRS classifies a specimen as positive if it is so by either or both of the reference and the resolver method, then the sensitivity of the test method is given by

$$\frac{P_{111} + P_{101} + P_{110}}{P_{+11} + P_{+01} + P_{+10}}$$

and the specificity by $P_{000}/P_{+00}$ Subsampling the overpopulated cells of the original 2×2 table will provide estimates of the 8 probabilities, and standard errors can be computed using the sampling variability of the estimated probabilities and the delta method, closely paralleling the earlier calculations.

Where the reference test has high sensitivity but only moderate specificity, it should be matched by a resolver with similar characteristics, and then a CRS can be set up defining as positive only those specimens for which both tests are positive, and as negative those for which either or both of the tests give a negative result. The calculations for this situation are formally identical to those of the CRS for a high specificity reference method.

For example, in Alonzo and Pepe's *Chlamydia* data set, taking culture and PCR to define a composite standard and taking EIA as a test method gives the 2×2×2 table

Table 8 – Alonzo and Pepe's data

| | | Culture | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Positive | | | Negative | | | |
| | | PCR | | | PCR | | | |
| | | Positive | Negative | Total | Positive | Negative | Total | Total |
| EIA | Positive | 20 | 0 | 20 | 4 | 3 | 7 | 27 |
| | Negative | 2 | 1 | 3 | 2 | 292 | 294 | 297 |
| | Total | 22 | 1 | 23 | 6 | 295 | 301 | 324 |

(The top left cell is a fiction. The 20 doubly positive specimens were not submitted to PCR; it was assumed that all would test positive by PCR). This gives the estimates

| | Sensitivity | Specificity | PPV | NPV |
|---|---|---|---|---|
| Estimate | 0.8276 | 0.9898 | 0.8889 | 0.9832 |
| SE | 0.0701 | 0.0058 | 0.0605 | 0.0075 |

Suppose instead of testing all 294 doubly negatives we were to test just 100, and found one of these positive. Then the estimates and standard errors would become

| | Sensitivity | Specificity | PPV | NPV |
|---|---|---|---|---|
| Estimate | 0.8016 | 0.9898 | 0.8889 | 0.9800 |
| SE | 0.0967 | 0.0059 | 0.0605 | 0.0113 |

The total number of samples submitted to PCR is reduced from 304 to 110; the precision of the specificity and PPV estimates deteriorates only slightly, though the standard error of the sensitivity and

NPV increase by some 40%. The subsampling is quite harmful to the precision of the sensitivity and NPV estimates (as also noted by Alonzo and Pepe), but not at all to the specificity or PPV estimates.

**Conclusion**

Evaluating the performance of a qualitative diagnostic test is not a trivial problem. Errors made by the reference but imperfect gold standard translate into artificially low estimates of the test's diagnostic performance. Using a perfect resolver method on the discrepant subjects at first sight gives one a way to cure these discrepancies, but on closer inspection this does not clear up, or necessarily even ameliorate, the problem. Proper resolution requires some testing of some subjects with concordant, as well as discordant, results.

This raises the possibility of subsampling, and applying the resolver to just some of the subjects. We derive estimators and standard errors for the resulting diagnostic performance figures. A numeric example illustrates that this approach allows precise estimates to be found with substantial saving in the number of resolver assays used. The subsampling approach is also effective when imperfect reference tests are combined in a composite reference standard.

**Appendix 1: Computation of *Var*(sensitivity) for the example data**

Turning to the original 2×2 table, the four cell frequencies follow a joint multinomial distribution, with the covariance between any two cells is given by

$$Cov(n_{ij+}, n_{km+}) = -nP_{ij+}P_{km+}$$

The resolver frequencies $r_{ij}$ involve separate tests of the four cells and can be expected to be statistically independent of each other. If we consider two generic terms involved in the estimates of true sensitivity and specificity, $p_{ij+}r_{ij}$ and $p_{km+}r_{km}$, their covariance is estimated by

$$Cov(p_{ij+}r_{ij}, p_{km+}r_{km}) = -p_{ij+}p_{km+}r_{ij}r_{km} / n \qquad (12)$$

Using these pairwise covariances, we can calculate the standard error of the estimated probabilities of true positives, of true negatives, and of their covariance. Likewise, the variances of prevalence and (1-prevalence) can be calculated. In general, the form of the variance of a sum of the $p_{ij+}r_{ij}$ is given by.

$$Var(\sum_i \sum_j p_{ij+}r_{ij}) = \sum_i \sum_j Var(p_{ij+}r_{ij}) + 2\sum_{i\le k} \sum_{j\le m} Cov(p_{ij+}r_{ij}, p_{km+}r_{km})$$

Specifically, the estimate of the variance of probability of true positive is given by:

$$\begin{aligned} Var(p_{11+}r_{11} + p_{10+}r_{10}) &= Var(p_{11+}r_{11}) + Var(p_{10+}r_{10}) + 2Cov(p_{11+}r_{11}, p_{10+}r_{10}) \\ &= (p_{11+})^2 \frac{r_{11}(1-r_{11})}{m_{11}} + (r_{11})^2 \frac{p_{11+}(1-p_{11+})}{n} + (p_{10+})^2 \frac{r_{10}(1-r_{10})}{m_{10}} \\ &+ (r_{10})^2 \frac{p_{10+}(1-p_{10+})}{n} - 2\frac{p_{11+}p_{10+}r_{11}r_{10}}{n} \end{aligned} \quad (13)$$

These terms are necessary for computing the standard errors for sensitivity and specificity that are used in generating confidence intervals for the estimates. Variances for sensitivity and specificity can be computed using the delta method for the variance of the ratio of two correlated variables (X,Y).

$$Var\left(\frac{X}{Y}\right) = \frac{Var(X)}{[E(Y)]^2} + \frac{[E(X)]^2}{[E(Y)]^4} Var(Y) - 2\frac{E(X)}{[E(Y)]^3} Cov(X,Y)$$

*The numeric example*

Recall (equation 11) that if $X = p_{11+}r_{11} + p_{10+}r_{10}$, the estimated probability of a true positive result by the index method, and $Y = p_{11+}r_{11} + p_{10+}r_{10} + p_{01+}r_{11} + p_{00+}r_{00}$, the estimated prevalence of the disease, then the estimated sensitivity $X/Y$ has a sampling variance given by equation (11).

In order to get an estimate of this variance, we need estimates of *Var(Y), Var(X)* and *Cov*(*X,Y)*, along with the proportion of true positives, *E*(*X*), and the estimated prevalence, *E*(*Y*).

The estimated proportion of true positive results by the index test was found to be 0.65. The estimated prevalence was found to be 0.6667. Equation 13 provides the form for calculating the variances of true positives and prevalence. Equations 7 and 12 give the forms for computing the variances and covariances of the $p_{ij+}r_{ij}$ terms.

Estimated standard errors for the joint probabilities $p_{ij+}r_{ij}$ are shown in Table A1:

Table A1: Estimated standard errors of $p_{ij+}r_{ij}$

| | | Reference test | |
|---|---|---|---|
| | $SE_{ij1}$ | Positive | Negative |
| Index test | Positive | 0.01559 | 0.01022 |
| | Negative | 0.00248 | 0.00292 |

Estimated covariances for the $p_{ij+}r_{ij}$, $p_{km+}r_{km}$ are given in Table A2:

Table A2: Estimated covariances

| | Term *i,j* | | |
|---|---|---|---|
| Term *k,m* | 1,1 | 1,0 | 0,1 |
| 1,0 | -1.520E-5 | | |
| 0,1 | -1.900E-6 | -2.667E-7 | |
| 0,0 | -1.267E-6 | -1.778E-7 | -2.222E-8 |

Using (13) and the corresponding standard error and covariance estimates from Tables A1 and A2, variances for the proportion of true positives and the prevalence can be computed. Note that for calculating the variance of prevalence, we need to compute the covariance of true positives and the prevalence. Since the prevalence is the sum of true positives and false negatives, the covariance term can be written as:

$$\begin{aligned} Cov(\text{True Positives}, \text{Prevalence}) &= Cov(\text{True Positives}, \text{True Positives} + \text{False Negatives}) \\ &= Var(\text{True Positives}) + Cov(\text{True Positives}, \text{False Negatives}) \end{aligned}$$

The various sub-calculations are not shown here. However, the following terms were calculated to match the format of Equation 13:

*SE*(True positives) = 0.01781
*SE*(Prevalence) = 0.01802
*Cov*(True positives, Prevalence) = 3.137E-4

Combining these results, the estimated standard error is found to be:

$$\begin{aligned} SE(Sensitivity) &= \sqrt{\frac{0.01781^2}{0.6667^2} + 0.65^2 \frac{0.01802^2}{0.6667^4} - 1.3\frac{3.137\times10^{-4}}{0.6667^3}} \\ &= \sqrt{7.1362\times10^{-4} + 6.9441\times10^{-4} - 1.37615\times10^{-3}} \\ &= 0.00565 \end{aligned}$$

**Bibliography**

Agresti, A., and Coull, B. A., (1998), "Approximate is better than "exact" for interval estimation if binomial proportions", *American Statistician* **52,** 119—126.

Alonzo, T.A., and Pepe, M. S., (1999), "Using a combination of reference tests to assess the accuracy of a new diagnostic test," *Statistics in Medicine* **18,** 2987-3003.

Begg, C. B. and Greenes, R. A. (1983), "Assessment of diagnostic tests when disease verification is subject to selection bias," *Biometrics*, **39**, 207-215.

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975), *Discrete Multivariate Analysis*, MIT Press, Cambridge.

Cochran, W. G., (1977), *Sampling Techniques, 3rd Edition*, Wiley, New York.

DeGroot, M. H. (1975), *Probability and Statistics*. Addison-Wesley Publishing Co., Reading, Massachusetts.

Fleiss, J. L. (1981), *Statistical Methods for Rates and Proportions*. John Wiley & Sons, New York.

Green, T. A., Black, C. M., and Johnson, R. E. (1998), "Evaluation of Bias in Diagnostic-Test Sensitivity and Specificity Estimates Computed by Discrepant Analysis," *Journal of Clinical Microbiology,* **36**:375–381.

Hadgu, A. and Qu, Y. (1998), "A Biomedical Application of Latent Class Models with Random Effects," *Applied Statistics,* Part 4, **47**:603–616.

Hadgu, A., (1996), "The discrepancy in discrepant analysis", *The 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.

Hui, S. and Walter, S. (1980), "Estimating the Error Rates of Diagnostic Tests," *Biometrics,* **36**:167–171.

Irwig, L., Glasziou, P. P., Berry, G., Chock, C., Mock, P., and Simpson, J. M., (1994), "Efficient study designs to assess the accuracy of screening tests", *American Journal of Epidemiology*, **140,** 759-769.

Joseph, L., Gyorkos, T., and Coupal, L. (1995), "Bayesian Estimation of Disease Prevalence and the Parameters of Diagnostic Tests in the Absence of a Gold Standard," *American Journal of Epidemiology,* **141**:263–272.

Meier, K. (1999), "Discrepant resolution for diagnostic devices", Talk presented to FDA/Industry Workshop "Statistical Issues for the New Millenium", Arlington VA, October 1.

Meier, K. (1998), FDA document, *Microbiology Devices Panel Medical Advisory Committee Meeting, Wednesday, February 11, 1998, 11:00 a.m.*, p. 25.

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

Qu, Y., Tan, M., Kutner, M. (1996), "Random Effects Models in Latent Class Analysis for Evaluating Accuracy of Diagnostic Tests," *Biometrics,* **52**:797–810.

Qu, Y. and Hadgu, A. (1998), "A Model for Evaluating Sensitivity and Specificity for Correlated Diagnostic Tests in Efficacy Studies with an Imperfect Reference Test," *Journal of the American Statistical Association,* 93:920–928.

Rindskopf, D., and Rindskopf, W. (1986), "The Value of Latent Class Analysis in Medical Diagnosis," *Statistics in Medicine,* **5**:21–27.

Shih, J. H. and Albert, P. A., (1999), "Latent model for correlated binary data with diagnostic error", *Biometrics*, **55**, 1232-1235.

Staquet, M., Rozencweig, M., Lee, Y., and Muggia, F. (1981), "Methodology for the Assessment of New Dichotomous Diagnostic Tests" *Journal of Chronic Diseases,* **34**:599–610.

Valenstein, Paul. (1990), "Evaluating Diagnostic Tests with Imperfect Standards," *American Journal of Clinical Pathology,* **93**:252–258.

Walter, S. D., and Irwig, L. M., (1988), "Estimation of test error rates, disease prevalence and relative risk from misclassified data: A review", *Journal of Clinical Epidemiology,* **41** 923-937.

Yang, S.-J. and Becker, M., (1997), "Latent variable modeling of diagnostic accuracy", *Biometrics***, 55,** 948—958.

*Revision, July 2002.*