Some Issues in Resolution using an Imperfect Gold Standard

Douglas M. Hawkins[1], James A. Garrett[2] and Betty Stephenson[3]

Technical Report #628

School of Statistics
University of Minnesota

March 1999

Some Issues in Resolution using an Imperfect Gold Standard

Douglas M. Hawkins[1], James A. Garrett[2] and Betty Stephenson[3]
Technical Report #628
School of Statistics
University of Minnesota
March 1999

Abstract

As the true status of a test specimen is seldom known with certainty, it is necessary to compare the performance of new diagnostic tests with those of a current accepted but imperfect "gold standard". Errors made by the gold standard mean that the apparent sensitivity and specificity of the new test relative to the true status are biased, and do not estimate the new method's "true" sensitivity and specificity. The traditional resolution of this problem was "discrepant resolution", in which the cases in which 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 would go beyond the discordant cases and test at least some concordant cases 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 have a simple universal answer. Another issue, if one is to test a sample of concordant cases rather than all cases, is that of 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 concordant cases may lead great savings in assays. The design issue of how many concordant cells to test depends on the numbers of concordant and discordant cases. The formulas given show the 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.

---

[1] Department of Applied Statistics, University of Minnesota, St. Paul, MN 55108
[2] Becton Dickinson
[3] Becton Dickinson

Introduction

A common problem faced in diagnostic testing is of determining the diagnostic performance of an "index test," that is, a test that yields a positive or negative result rather than a quantitative measurement. The performance of such a test is usually summarized by its sensitivity and specificity; sensitivity being the proportion of truly positive samples that yield positive test results, and specificity the proportion of truly negative samples that yield negative test results.

In an ideal world, one can estimate sensitivity and specificity by applying the index test to a series of samples whose true disease status is known with certainty. Estimation of sensitivity and specificity is straightforward in this case; and likewise the associated confidence intervals are easy to calculate, see for example Fleiss, 1973. It is more often the case, however that perfect diagnosis of the samples is impossible, unethical, or impracticably expensive to obtain and one must settle for 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 a hypothetical 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 classical "discrepant analysis," or "discrepant resolution," whereby samples yielding different test results by the index and reference method are retested by a third "resolving" method. Variations on this design have been discussed such as retesting only apparently false negatives or only apparently false positives. While discrepant analysis of any form is intended to yield additional information about potentially problematic samples, it introduces its own set of biases, always in an upward direction. There has been some debate concerning whether these biases more or less counteract misclassification biases, and whether discrepant resolution is the lesser of two evils (see Hadgu, 1997; and Green, 1998). In this paper, we 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 samples. The literature on one such method, Latent Class analysis (LCA) is large, and some 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 used mathematical adjustments using prior knowledge. All of these efforts assume that index and reference tests err independently, a highly questionable assumption in most cases. Recently Qu et. al (1996), Qu and Hadgu (1998), and Hadgu and Qu (1998) 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 concordant samples in addition to the discrepant samples. This idea is appealing, however no method of statistical analysis for this design has been documented. We begin to fill this void by proposing a method for estimating sensitivity and specificity, and for calculating associated approximate confidence intervals. This method does not assume conditional independence, nor is it based

on a latent class model. We apply the method to a hypothetical example, and find that this study design, analyzed with our proposed method, makes efficient use of resolving tests.

Estimating Sensitivity and Specificity When Sample Status Is Known

Suppose an index test is applied to a group of subjects whose true disease status was known. Setting up the 2 × 2 cross-classification of the subjects test result 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.

In symbols, write the 'true' classification table

Table 1:

	True status		
Index test	D	N	Total
Positive	n_{1D}	n_{1N}	n_{1+}
Negative	n_{2D}	n_{2N}	n_{2+}
Total	n_{+D}	n_{+N}	n_{++}

The estimated sensitivity and its estimated standard error are given by

$$Sens = n_{1D} / n_{+D}, \quad se_{sens} = \sqrt{\frac{n_{1D} n_{2D}}{n_{+D}^3}} \tag{1}$$

and the corresponding formulas for specificity are

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

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

In both cases, one is forced to use an imperfect gold standard as the reference. This contaminates the 'true table', giving the observed table.

Table 2: Observed Table

	Reference test		
Index test	Positive	Negative	Total
Positive	n_{11}	n_{12}	n_{1+}
Negative	n_{21}	n_{22}	n_{2+}
Total	n_{+1}	n_{+2}	n_{++}

The 'apparent' or 'relative' sensitivity and specificity of the index test are

$$\textit{Apparent sensitivity} = n_{11} / n_{+1}$$
$$\textit{Apparent specificity} = n_{22} / n_{+2}$$

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 – even as to the direction of the bias. The difficulty in determining the potential biases can be illustrated by subdividing the true table's columns by the reference diagnosis as shown in Table 3.

Table 3:

		Truth						
		D			N			
		Reference			Reference			
		Positive	Negative		Positive	Negative		Total
Test	Positive	n_{1D}-a	a	n_{1D}	c	n_{1N}-c	n_{1N}	n_{1+}
	Negative	n_{2D}-b	b	n_{2D}	d	n_{2N}-d	n_{2N}	n_{2+}
	Total	n_{+D}-(a+b)	a+b	n_{+D}	c+d	n_{+N}-(c+d)	n_{+N}	n_{++}

The letters *a*, *b*, *c* and *d* represent errors by the reference method; *a* and *b* are false negatives, *c* and *d* are false positives. Rearranging the table according to true diagnosis within reference result gives the following table.

Table 4:

		Reference						
		Positive			Negative			
		Truth			Truth			
		D	N	Total	D	N	Total	Total
Test	Positive	n_{1D}-a	c	n_{1D}-a+c	a	n_{1N}-c	n_{1N}+a-c	n_{1+}
	Negative	n_{2D}-b	d	n_{2D}-b+d	b	n_{2N}-d	n_{2N}+b-d	n_{2+}
	Total	n_{+D}-(a+b)	c+d	n_{+D}-a+c-b+d	a+b	n_{+N}-(c+d)	n_{+N}+a-c+b-d	n_{++}

Collapsing over the true disease status results in this observed table as shown in Table 5.

Table 5:

		Reference		Total
		Positive	Negative	
Test	Positive	$n_{11}=n_{1D}$-a+c	$n_{12}=n_{1N}$+a-c	n_{1+}
	Negative	$n_{21}=n_{2D}$-b+d	$n_{22}=n_{2N}$+b-d	n_{2+}
	Total	$n_{+1}=n_{+D}$-(a+b)+(c+d)	$n_{+2}=n_{+N}$+(a+b)-(c+d)	n_{++}

Using the results from Table 5, the apparent sensitivity and specificity of the index test can be written

$$\textit{Apparent sensitivity} = (n_{1D}-a+c) / [n_{+D}-a+c-b+d] \quad (3)$$
$$\textit{Apparent specificity} = (n_{2N}+b-d) / [n_{+N}+a-c+b-d]$$

These expressions show that the connections between the true and the apparent sensitivity and specificity are not straightforward. Depending on how errors in the reference method distribute among the four cells, the apparent sensitivity and specificity may be biased either upward or downward. It is also possible for the net errors *c-a*and *b-d* to be zero, resulting in unbiased apparent sensitivity and specificity estimates. Sweeping generalizations about the bias in the original relative sensitivity and specificity, therefore, require considerable caution.

Using a resolver in classical 'discrepant analysis'.

The presence of bias in the apparent sensitivity and specificity leads to the idea of using a further test to locate and correct the errors in the reference method. In its classic form, 'discrepant analysis' tests the cells where the index was positive and the reference negative and uses this to calculate a 'resolved' sensitivity. A similar resolution of the cases where the test as negative and the reference positive is used in the same way to get a 'resolved' specificity.

As has been pointed out recently, this traditional approach to discrepant analysis does not remove all biases in the apparent sensitivity and specificity. If a perfectly specific test is used to resolve the cases in the top right cell, only the a cases in which the reference method gave a false negative will be corrected. It will not identify the c cases in which the reference method gave a false positive, or the $b+d$ cases where the index test was falsely negative. Because the bias involves all four quantities a, b, c and d, 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 on the lower left cell of the table does not lead to an unbiased estimate of specificity.

While classical discrepant analysis leads to 'resolved' sensitivity and specificity estimates that are always higher than the apparent sensitivity and specificity, it is not automatically true that it is more biased than the apparent 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 an exact resolver to produce unbiased sensitivity and specificity estimates

Classical discrepancy 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, at a minimum, the differences *(c-q)* and *(b-d)*. These can be found most 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. 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. Use the value '1' for i, j and k to indicate as positive test, or a diseased status; and the value '2' to indicate 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 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_{2+2}/P_{++2}.

Estimates of the 8 probabilities in the 2×2×2 table can be found by taking n subjects and testing them using the index and reference tests and also the exact resolver. Write n_{ijk} for the number of subjects in the i, j, k cell of the table. Estimates of p_{ijk} and its corresponding 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 estimates and standard errors of the true sensitivity and specificity by margining out the reference test.

Sampling approaches

This whole scenario as detailed above, however is unrealistic. If the exact resolver could reasonably be used on all cases, it would likely be used as the reference method. Often it is not used is because it is difficult or expensive. In this case, it cannot be used on all n cases but it could reasonably be used on a subset of the cases. This raises the possibility of subsampling some or all of the cells of the 2×2 table and evaluating those cases with the exact resolver.

Returning to the collapsed 2×2 table defined by the index and reference tests, the observed frequencies are:

		Reference Test		
		Positive	Negative	Total
Index Test	Positive	n_{11+}	n_{12+}	n_{1++}
	Negative	n_{21+}	n_{22+}	n_{2++}
	Total	n_{+1+}	n_{+2+}	n

The estimates of the 2×2 marginal probabilities P_{ij+} and standard errors are give 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 exact 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 standard error is given by

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

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_{ijl}. The standard error of this estimate can be found approximately using the 'delta' method by

$$Var(XY) = [E(X)]^2 Var(Y) + [E(Y)]^2 Var(X)$$

as (7)

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

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

$$p_{11+}r_{11} + p_{12+}r_{12} \quad (8)$$

The probability P_{2+1} of the index method giving a false negative is estimated as

$$p_{21+}r_{11} + p_{22+}r_{22} \quad (9)$$

The true sensitivity can then be estimated by

$$(p_{11+}r_{11} + p_{12+}r_{12})/(p_{11+}r_{11} + P_{12+}r_{12} + p_{21+}r_{11} + P_{22+}r_{22}) \quad (10)$$

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. 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

$$Cov(n_{ij+}, n_{km+}) = -nP_{ij+}P_{km+} \quad (11)$$

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

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

Using these pairwise covariances, we can calculate the standard error of the estimates of true positives, that of true negatives, and 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 Equation 13.

$$Var(\sum_i \sum_j p_{ij+}r_{ij}) = \sum_i \sum_j Var(p_{ij+}r_{ij}) + 2\sum_{i \le k} \sum_{j \le m} Cov(p_{ij+}r_{ij}, p_{km+}r_{km}) \quad (13)$$

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

$$\begin{aligned} Var(p_{11}r_{11} + p_{12}r_{12}) &= Var(p_{11}r_{11}) + Var(p_{12}r_{12}) + 2Cov(p_{11}r_{11}, p_{12}r_{12}) \\ &= (p_{11+})^2 \frac{r_{11}(1-r_{11})}{m_{11}} + (r_{11})^2 \frac{p_{11}(1-p_{11})}{m_{11}} + (p_{12+})^2 \frac{r_{12}(1-r_{12})}{m_{12}} \\ &+ (r_{12})^2 \frac{p_{12}(1-p_{12})}{m_{12}} - 2\frac{p_{11+}p_{12+}r_{11}r_{12}}{n} \end{aligned} \quad (14)$$

These terms are necessary for computing that 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 by using the delta method to generate an expression for the variance of the ratio of two correlated variables (A,B).

$$Var\left(\frac{A}{B}\right) = \frac{Var(A)}{[E(B)]^2} + \frac{[E(A)]^2}{[E(B)]^4}Var(B) - 2\frac{E(A)}{[E(B)]^3}Cov(A,B) \quad (15)$$

In terms of sensitivity, this becomes:

$$\begin{aligned} Var\left(\frac{\text{True Positives}}{\text{Prevalence}}\right) &= \frac{Var(\text{True Positives})}{[E(\text{Prevalence})]^2} + \frac{[E(\text{True Positives})]^2}{[E(\text{Prevalence})]^4}Var(\text{Prevalence}) - \\ &2\frac{E(\text{True Positives})}{[E(\text{Prevalence})]^3}Cov(\text{True Positives, Prevalence}) \end{aligned} \quad (16)$$

A (1-α)*100% confidence interval can be generated for sensitivity (specificity) using a normal approximation.

Example

A numeric example may help to clarify the forest of symbols. The following numbers are not from any actual data set, but are chosen to illustrate the methods of calculation. Consider a condition with a prevalence of 65% and classified by an index test and an imperfect resolver, leading to the data

		Reference test		
	n_{ij+}	Positive	Negative	Total
Index test	Positive	1800	600	2400
	Negative	200	400	600
	Total	2000	1000	3000

yielding the table of proportions

Table 7:

		Reference test	
	p_{ij+}	Positive	Negative
Index test	Positive	0.600	0.200
	Negative	0.067	0.133

Suppose we decide to test 200 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 8:

Resolver Test Positive			
		Reference test	
	$m_{ij}(r_{ij})$	Positive	Negative
Index test	Positive	190 (.95)	80 (.40)
	Negative	30 (.15)	30 (.15)

Using the probabilities computed in Tables 7 and 8, the joint probabilities ($p_{ijl} = p_{ij+}\, r_{ij}$) can be computed. This leads to the estimates of the probabilities of the full 2×2×2 table. We show only the $ij+$ cells in Table 9 – those in which the truth is that the subject is diseased since $P_{ij2} = P_{ij+} - P_{ijl}$, the omitted cells (true negatives) can be found by simple subtraction of Table 9 from Table 7.

Table 9:

		Reference test	
	$p_{ij}r_{ij}$	Positive	Negative
Index test	Positive	0.570	0.080
	Negative	0.010	0.020

Using the in Table 9, the following parameters can be estimated:

True positive by the index test	0.57 + 0.08 = 0.65.
False negative by the index test	0.01 + 0.02 = 0.03.
Prevalence	0.65 + 0.03 = 0.68
True sensitivity of the index test	0.65/0.68 = 0.956, or 95.6%.

Note the difference between the estimated true sensitivity of the index test as computed using the perfect resolver information and the sensitivity of the index test relative to the [imperfect] reference test of 90% (1800/2000*100). In order to calculate a confidence interval for sensitivity, we need to calculate the standard error based on (15). The intermediate calculations for the standard error are shown in Appendix 1. The calculated value is found to be 0.007. A 95% confidence interval for sensitivity (using a normal approximate) is:

$$\text{Sensitivity} = 0.955 \pm 2*0.07 = 0.955 \pm 0.014$$

This confidence interval width matches what could be found from a sample of 940 true positive specimens assayed by the index test. With a prevalence of 65%, if we had validated the index test directly

against the resolver, to attain this precision would have required n=1450. The actual number of resolver assays used, 800, is little more than half this number. This represents a substantial saving in resolver assays. If the resolver were a 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.

Approaches where there is no exact resolver

In the majority of applications, there is no exact resolver. There may however be a third and perhaps fourth and further test(s) that could be applied in an effort to improve the picture of the true status of the subjects.

The same approach of sampling can be followed to produce estimates of the 2×2×2 table. Here however the use of the table is less clearcut than it was in the case where the third test provided an unambiguously correct classification. Possible uses of the table include the fitting of latent class models, and more complex 'voting' rules involving multiple reference tests. Latent class models would require an accounting for repeated testing, as well as a provision for conditional dependence of assay errors. These further steps lie outside the scope of this note.

Summary

Evaluating the performance of a qualitative index 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 resolver method on the discrepant cases 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 concordant, as well as discordant, samples.

This raises the possibility of subsampling, and applying the resolver to just some of the samples. 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.

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

As shown in the text, the expression for the variance of the sensitivity is given by:

$$Var\left(\frac{\text{True Positives}}{\text{Prevalence}}\right) = \frac{Var(\text{True Positives})}{[E(\text{Prevalence})]^2} + \frac{[E(\text{True Positives})]^2}{[E(\text{Prevalence})]^4} Var(\text{Prevalence}) - 2\frac{E(\text{True Positives})}{[E(\text{Prevalence})]^3} Cov(\text{True Positives, Prevalence})$$

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

The estimated proportion of true positive results by the index test was found to be 0.65. The estimated prevalence was found to be 0.68. 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:

		Reference test	
	se_{ij+}	Positive	Negative
Index test	Positive	0.0126	0.0075
	Negative	0.0018	0.0035

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

Table A2:

	Term *i,j*		
Term *k,m*	1,1	1,2	2,1
1,2	-1.52E-5		
2,1	-1.9E-6	-2.7E-7	
2,2	-3.8E-6	-5.3E-7	6.7E-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:

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

The various sub-calculations are not shown here, however, the following terms were calculated to match the format of Equation 16:

$Var(\text{True positives}) = 0.014$
$Var(\text{Prevalence}) = 0.0136$
$Cov(\text{True positives, Prevalence}) = 1.8\text{E-}4$

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

$$se(\text{Sensitivity}) = \sqrt{\frac{0.014^2}{0.68^2} + 0.65^2\frac{0.0136^2}{0.68^4} - 1.3\frac{1.8\text{E}-4}{0.68^3}} = 0.007$$

Bibliography

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. (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.

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.

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.

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.

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