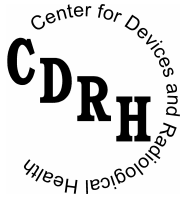
Statistical Guidance on Reporting Results from Studies Evaluating Diagnostic Tests; Draft Guidance for Industry and FDA Reviewers

Draft Guidance – Not for Implementation

This guidance document is being distributed for comment purposes only.
Draft released for comment on March 12, 2003



U.S. Department of Health and Human Services
Food and Drug Administration
Center for Devices and Radiological Health

Division of Biostatistics
Office of Surveillance and Biometrics

Preface

Public Comment:

For 90 days following the date of publication in the Federal Register of the notice announcing the availability of this guidance, comments and suggestions regarding this document should be submitted to the Docket No. assigned to that notice, Dockets Management Branch, Division of Management Systems and Policy, Office of Human Resources and Management Services, Food and Drug Administration, 5630 Fishers Lane, Room 1061, (HFA-305), Rockville, MD 20852.

Additional Copies

Additional copies are available from the Internet at: http:www.fda.gov/cdrh/osb/guidance/1428.pdf, or CDRH Facts-On-Demand. In order to receive this document via your fax machine, call the CDRH Facts-On-Demand system at 800-899-0381 or 301-827-0111 from a touch-tone telephone. Press 1 to enter the system. At the second voice prompt, press 1 to order a document. Enter the document number (1428) followed by the pound sign (#). Follow the remaining voice prompts to complete your request.

Statistical Guidance on Reporting Results from Studies Evaluating Diagnostic Tests; Draft Guidance for Industry and FDA Reviewers

This document is intended to provide guidance. It represents the Agency's current thinking on this topic. It does not create or confer any rights for or on any person and does not operate to bind the Food and Drug Administration (FDA) or the public. An alternative approach may be used if such approach satisfies the requirements of the applicable statute and regulations.

Background

On February 11, 1998, the Center for Devices and Radiological Health convened a joint meeting of the Microbiology, Hematology/Pathology, Clinical Chemistry/Toxicology and Immunology Devices Panels. The purpose of this meeting was to obtain recommendations on "appropriate data collection, analysis, and resolution of discrepant results, using sound scientific and statistical analysis to support indications for use of the *in vitro* diagnostic devices when the new device is compared to another device, a recognized reference method or 'gold standard', or other procedures not commonly used, and/or clinical criteria for diagnosis." Using the input from that meeting, this document discusses some statistically valid approaches to reporting results from evaluation studies for new diagnostic devices.

Scope

This document provides guidance for the submission of premarket notification (510(k)) and premarket approval (PMA) applications for diagnostic tests. This guidance addresses the reporting of results from different types of studies evaluating diagnostic devices with two possible outcomes (positive or negative).

This guidance does not address the fundamental statistical issues associated with design and monitoring such clinical studies.

Purpose

This guidance is intended to describe some statistically appropriate practices for reporting results from different studies evaluating diagnostic tests, and to identify common inappropriate practices. Special attention is given to describing a practice called *discrepant resolution* and its associated problems.

Introduction

A diagnostic test is a measurement used to indicate the presence or absence of a specific disease or condition in a patient from a specific patient population. In the simplest case, a patient either has the disease or condition or does not have the disease or condition. This document uses the terms diseased and non-diseased generically to refer to a disease or condition of interest. A qualitative diagnostic test indicates whether the patient is diseased or non-diseased. For this simple case, two quantities Jointly measure test performance: (clinical) sensitivity and (clinical) specificity. Sensitivity is how often the test is right in diseased patients, and specificity is how often the test is right in non-diseased patients. (Note that *clinical* sensitivity and specificity are different from *analytical* sensitivity and specificity. Analytical sensitivity measures a test's ability to detect a low concentration of a given substance, and analytical specificity measures a test's ability to exclusively identify a target substance rather than similar but different substances).

More specifically,

- *Clinical Sensitivity* is how often the test is positive in diseased patients
- *Clinical Specificity* is how often the test is negative in non-diseased patients.

There are different ways to evaluate a new diagnostic test. One aspect of evaluating a new diagnostic test involves testing specimens from patients who are representative of the target patient population, and comparing the outcome of a new test with the clinical status of a patient or with the outcome of some other procedure. When the comparative procedure is generally accepted as an indicator of true clinical status by the clinical community and is regarded as having negligible risk of having either a false positive or false negative result, we say that the comparative procedure is a "perfect standard" for performance estimation purposes. The standard should be definitive (positive/negative, present/absent, diseased/non-diseased) and should not give an indeterminate result.

When the new test is compared to clinical status or to a perfect standard, the sensitivity of the new test is estimated as the proportion of specimens from diseased patients where the test is positive. Similarly, the specificity of the test is estimated as the proportion of specimens from non-diseased patients where the test is negative (See the Appendix for an example of this calculation). These are only *estimates* for sensitivity and specificity because they are based on only a subset (sample) of specimens; if another subset of specimens were tested, the estimates of sensitivity and specificity

would probably be numerically different. However, if the specimens are *representative* of specimens from the target patient population, the estimates will be statistically *unbiased* (on average, the estimates will equal the true sensitivity and specificity).

Knowing whether the patient is truly diseased or non-diseased is a critical issue in estimating sensitivity and specificity. If the comparative procedure is imperfect, then sensitivity and specificity estimates are almost always statistically biased (*systematically* too high or too low). Even worse, the size and direction of the bias usually cannot be determined; the only thing that can be said is that the estimates will be inaccurate due to this bias. So, to obtain unbiased estimates of sensitivity and specificity, the new device should be compared to true clinical status or to a perfect standard.

Sometimes, however, comparing a new test to clinical status or to a perfect standard is impossible, impractical, or extremely expensive. Instead, new tests are often evaluated by comparison to an imperfect standard. In this situation, sensitivity and specificity are not appropriate terms to describe the comparative results. The question is how to report results from a study evaluating a new diagnostic test when the comparative procedure is imperfect.

This document suggests some statistically appropriate practices for reporting results from different studies evaluating a new diagnostic test under these circumstances, and identifies some common practices that are statistically inappropriate.

General Statistical Guidance for Evaluating a New Diagnostic Test

The most important advice is to carefully plan the study before collecting the first specimen. This includes determining whether you want to report sensitivity and specificity. If you want to report these measures, then your evaluation needs to include using the patients' clinical status or a perfect standard on at least some of the specimens.

Another key step in planning may be contacting CDRH to discuss possible study designs and statistical analyses prior to any data collection for the clinical study. There are promising advanced statistical methods that may be appropriate, and new statistical analysis techniques are constantly being developed. The list of references at the end of this document includes a variety of approaches. Discussing your planned study with CDRH before starting may save time and money.

In addition to careful planning, here are four general recommendations regarding choosing a comparative procedure to evaluate a new diagnostic test and reporting the results.

1. If a perfect standard is available, use it. Calculate estimated sensitivity and specificity.
2. If a perfect standard is available but impractical, use it to the extent possible. Calculate adjusted estimates of sensitivity and specificity.
3. If a perfect standard is not available, consider constructing one. Calculate estimated sensitivity and specificity under the constructed standard.
4. If a perfect standard is not available and cannot be constructed, then an appropriate approach may be reporting a measure of *agreement* (see Appendix).

These recommendations are now described in more detail.

- **If a perfect standard is available, use it.**

From a purely statistical perspective, the best approach is to compare the new test to the patients' clinical status or to a perfect standard using specimens from patients who are representative of the intended use population. In this situation, sensitivity and specificity have meaning and you can easily calculate the estimates. The Appendix describes a numerical example.

- **If a perfect standard is available but impractical, use it to the extent possible.**

If using a perfect standard across-the-board is considered impractical or not feasible, you could still obtain estimates of sensitivity and specificity if you use the new test and an imperfect standard on all specimens, and use the perfect standard on just a subset of specimens. For example, you could apply the perfect standard to all specimens where the new test and the imperfect standard disagree and to a random sample of specimens where they agree. Using these results, you can compute adjusted estimates (and variances) of sensitivity and specificity. However, you still may need to retest a large number of specimens in order to estimate sensitivity and specificity with reasonable precision. (In some instances, the required subset may be so large that testing all the specimens with the perfect standard in the first place may actually be simpler.) Since this approach can be statistically complicated, FDA recommends that you consult with a CDRH statistician before using this approach.

In rare circumstances, it may be possible to estimate sensitivity and specificity without using a perfect standard in the study. This may be reasonable when the sensitivity and specificity of the imperfect standard are well established from previous evaluations against the perfect standard on similar patient populations. Using the sensitivity and specificity of the imperfect standard, mathematical adjustments can be made to the comparative results between the new test and the imperfect standard to obtain statistically unbiased sensitivity and specificity estimates for the new test. However, all mathematical adjustments are based on assumptions that need to be verified. For example, one commonly used adjustment assumes that the new test and the imperfect standard are *conditionally independent* (that is, measure unrelated things) given the true state of disease. This assumption may not be appropriate since the new test is often designed to measure the same analyte (or a closely related substance) as the existing imperfect test.

- **If a perfect standard is not available, consider constructing one.**

Perhaps an expert panel can develop a set of clinical criteria (or a combination of reference tests and confirmatory clinical information) that would serve as a "perfect standard." While this approach may be more time consuming up front, if successful, you can easily calculate estimates of sensitivity and specificity. It is important that the test label clearly describe the standard that was constructed.

- **If a perfect standard is not available and cannot be constructed, then an appropriate approach may be reporting agreement.**

When a new test is evaluated by comparison to an imperfect standard, you cannot directly calculate unbiased estimates of sensitivity and specificity. Therefore, the terms sensitivity and specificity are

not appropriate to describe the comparative results. Instead, you should report the 2×2 table of results comparing the new test with the imperfect standard, a description of the imperfect standard, and a measure of agreement and its confidence interval. The Appendix describes a numerical example.

There are two major disadvantages with agreement measures. First, agreement is not a measure of "correctness" because two tests could agree on an incorrect diagnosis. In fact, both tests could agree well but both have poor sensitivity and specificity. Second, agreement results are dependent on the *disease prevalence* in the study patient population (number of diseased patients divided by the total number of diseased and non-diseased patients in the study). That is, different patient populations with different disease prevalence (typically unknown in this case) will almost always have different (sometimes substantially) agreement.

As a hypothetical example, suppose that the new test and the imperfect standard agree closely when the true diagnosis is negative, but they do not agree very well when the true diagnosis is positive. The overall percent agreement between the two methods will be higher in a patient population with low disease prevalence, and lower in a patient population with high disease prevalence. The performance (correctness) of the new test relative to the imperfect standard does not change, but the agreement changes because disease prevalence changes. If the disease prevalence is unknown, then it is unclear how to generalize the agreement measure to another patient population.

General Reporting Recommendations

In addition to choosing an appropriate comparative procedure, evaluating a new test also involves choosing appropriate patients, specimens, and individuals performing the tests. Further discussion of these design issues is beyond the scope of this document. However, all descriptions of comparative results should include a clear description of all methods used, and how and what data were collected. This includes:

- patient recruitment procedures,
- patient demographics,
- patient and specimen inclusion/exclusion criteria,
- specimen collection procedures,
- time of specimen collection and testing,
- types of specimens collected,
- number of specimens collected and tested and number discarded
- number of specimens included in final data analysis,
- specimen collection devices (if applicable), and
- specimen storage and handling procedures.

The data description should include an accounting of all patients and test results (number of tests planned, tested, discarded, used in final analysis) and descriptive summaries of the final results. Results should be reported overall, by site, and by any other relevant categories.

For qualitative tests derived from an underlying quantitative result, descriptive summaries should include ranges of results, histograms of results by disease state (if known), and Receiver Operating Characteristic (ROC) Plots (if disease state is known). The most current edition of NCCLS document GP10 - Assessment of the Clinical Accuracy of Laboratory Tests Using Receiver Operating Characteristic (ROC) Plots provides further guidance on this topic.

All statistical measures such as sensitivity, specificity, and agreement should be reported both as fractions (e.g., 490/500) *and* as percentages (e.g., 98.0%). They should also be accompanied by confidence intervals to reflect the precision of the statistical measure. All statistical models and corresponding assumptions used to analyze the data should be clearly described.

Common Reporting Practices that are Statistically Inappropriate

Some common practices for reporting results are statistically inappropriate because they are misleading or can lead to inaccurate estimates of test performance. These practices can arise when a new test is compared to an imperfect standard.

Comparing a new test to an imperfect standard does not give true performance. If the new test is better than the imperfect standard, then the new test will not agree with the imperfect standard and the agreement will be poor. Alternatively, the agreement could be poor because the imperfect standard is fairly accurate and the new test is inaccurate. Which scenario is the true situation? There isn't any simple statistical solution to this dilemma.

However, when comparing a new test to an imperfect standard, there are three common practices listed below that should not be used because they give misleading or incorrect information.

1. Using the terms "sensitivity" and "specificity" to describe the comparison of a new test to an imperfect standard.
2. Using results from discrepant resolution alone to estimate the sensitivity and specificity of a new test or agreement between a new test and a comparative method.
3. Comparing the results of a new test to the outcome of a testing algorithm that combines several comparative methods, if the algorithm uses the outcome of the new test.

The problems with each of these practices and possible alternatives are discussed below.

- **You should not use the terms "sensitivity" and "specificity" to describe the comparison of a new test to an imperfect standard is inappropriate.**

When a new test is evaluated by comparison to an imperfect standard, it is impossible to directly calculate unbiased estimates of sensitivity and specificity. For this reason, FDA recommends that you report the 2×2 table of results comparing the new test with the imperfect standard, a description of the imperfect standard, and a measure of agreement and its confidence interval. (See Appendix).

- **You should not use results from discrepant resolution alone to estimate the sensitivity and specificity of a new test or agreement between a new test and a comparative method is inappropriate.**

When a new test is evaluated by comparison to an imperfect standard, discrepancies (disagreement) between the two methods may arise due to errors in the test method or due to errors in the imperfect standard. Since the imperfect standard may be wrong, calculations of sensitivity and specificity based on the imperfect standard are statistically biased. A practice called discrepant resolution has been suggested to get around the bias problem.

As the name implies, discrepant resolution focuses on specimens where the new test and the imperfect standard disagree. In the simplest situation, discrepant resolution can be described as a two-stage testing process in the following manner.

- Stage 1 - Test all specimens using the new test and the imperfect standard
- Stage 2 - When the new test and imperfect standard disagree, use an additional test (perfect standard) to see which one is "right"

A numerical example describing discrepant resolution is in the Appendix. While this process provides the true disease state for the re-tested specimens, it does not provide the true disease state for specimens when the new test agrees with the imperfect standard (usually most of the specimens). Even when the new test and imperfect standard agree, they may both be wrong.

It has been the practice of some to use the resolver results to revise the original 2×2 table of results (new test versus imperfect standard). The original 2×2 table is "revised" using the following reasoning.

- When the original two results agree, assume (without supporting evidence) that they are both correct and do not make any changes to the table.
- When the original results disagree and the imperfect standard disagrees with the resolver, assume that the resolver is correct and reclassify (change) the imperfect standard result to the resolver result.

The revised 2×2 table based on discrepant resolution is misleading because the columns are not clearly defined and do not necessarily represent truth, as assumed. The assumption that results that agree are correct isn't tested and may be far from valid. Such a table should not be presented because it may be very misleading. In addition, the calculation of sensitivity and specificity from a revised 2×2 table are not valid estimates of performance and should not be presented.

There do not seem to be any scientifically valid ways to estimate sensitivity and specificity by resolving only the discrepant results (unless all the specimens are discrepant and tested by the perfect standard!) even though the resolver is a perfect standard.

Discrepant resolution with a perfect standard can tell you whether the new test or the imperfect standard is right more of the time, but you can't quantify how much more. If the resolver is not a perfect standard, then the resolver test results do not provide useful information about the performance of the new test. Resolving discrepancies using repeat testing by the new test or the imperfect standard doesn't provide any useful information about performance either.

- **You should not compare the results of a new test to the outcome of a testing algorithm that combines several reference methods, if the algorithm uses the outcome of the new test.**

When evaluating some types of tests, the comparative "procedure" is not a single test but the outcome of a combination of several reference methods and possibly clinical information. Often, two or more reference methods are performed and interpreted according to a pre-specified testing sequence or algorithm to determine disease status. The decision to use a second or third reference method may depend on the outcome of the initial reference method. This approach may be statistically reasonable. However, this approach is not valid if the algorithm uses the outcome of the new unproven test. For example, the decision to use an additional reference method should not be based on whether the new test is positive or negative.

It is unscientific and potentially very misleading to establish the performance of a new test by comparing it to a procedure that uses the same new test.

APPENDIX

Calculating Estimates of Sensitivity and Specificity

Sensitivity and specificity are basic measures of performance for a diagnostic test. Together, they describe how well a test can determine whether a specific disease, or condition is present or absent. They each provide distinct and equally important information, and should be presented together.

- *Sensitivity* is how often the test is positive when the disease is present.
- *Specificity* is how often the test is negative when the disease is absent.

Usually, to estimate sensitivity and specificity, the outcome of the new test is compared to the true diagnosis using specimens from patients who are representative of the intended use (both diseased and non-diseased) populations. Results are typically reported in a 2×2 table such as Table 1.

Table 1. Common 2×2 table format for reporting results comparing a new test to true diagnosis. The new test has 2 possible outcomes, positive (+) or negative (–). Diseased patients are indicated as positive (+) true diagnosis, and non-diseased patients are indicated as negative (–) true diagnosis.

		True Diagnosis	
		+	–
New	+	a	b
Test	–	c	d
Total		a+c	b+d

From Table 1, estimated sensitivity is the proportion of diseased individuals that are New Test+. Estimated specificity is the proportion of non-diseased individuals that are New Test–. The formulas are as follows.

estimated sensitivity = 100% ´ a/(a+c)
estimated specificity = 100% ´ d/(b +d)

Here is an example of this calculation. Suppose one specimen is taken from each of 220 patients in the intended use population. Each specimen is tested by the new test, and the diagnosis for each patient is determined. Fifty-one (51) patients have the disease and 169 do not. The results are presented in a 2×2 table format in Table 2.

Table 2. Example of results comparing a new test to true diagnosis for 220 patients.

		True Diagnosis		
		+	–	Total
New	+	44	1	45
Test	–	7	168	175
Total		51	169	220

From Table 2, estimated sensitivity and specificity are calculated in the following manner.

$$\text{estimated sensitivity} = 100\% \times 44/51 = 86.3\%$$
$$\text{estimated specificity} = 100\% \times 168/169 = 99.4\%$$

Exact 95% confidence intervals (based on the binomial distribution) for sensitivity and specificity are (73.7%, 94.3%) and (96.8%, 100%), respectively.

Other quantities can be computed from this 2×2 table, too. These include positive predictive value, negative predictive value, and the positive and negative likelihood ratios. These quantities provide useful insight into how to interpret test results. However, further discussion of these quantities is beyond the scope of this document.

Calculating an Estimate of Agreement

When a new test is compared to an imperfect standard rather than to clinical diagnosis or to a perfect standard, the usual calculations from the 2×2 table, a/(a+c) and d/(b+d), respectively, are biased estimates of sensitivity and specificity because the imperfect standard is not always correct. In addition, quantities such as positive predictive value, negative predictive value, and the positive and negative likelihood ratios cannot be computed since truth is unknown. However, being able to describe how often a new test agrees with an imperfect standard may be useful. To do this, a group of individuals (or specimens from individuals) is tested twice, once with the new test and once with the imperfect standard. The results are compared and can be reported in a 2×2 table such as Table 3.

Table 3. Common 2×2 table format for reporting results comparing a new test to an imperfect standard.

		Imperfect Standard	
		+	–
New	+	a	b
Test	–	c	d
Total		a+c	b+d

The difference between Table 3 and Table 1 is that the columns of Table 3 do not represent truth, so data from Table 3 cannot be interpreted in the same way as Table 1. Data from Table 1 provides information on how often the new test is correct, whereas data from Table 3 provides information on how often the new test agrees with an imperfect standard.

From Table 3, you can compute several different statistical measures of agreement. A discussion by M.M. Shoukri on different types of agreement measures can be found under "Agreement, Measurement of" in the *Encyclopedia of Biostatistics*. Two commonly used measures are the overall percent agreement and Cohen's kappa. The simplest measure is overall percent agreement. Overall percent agreement is the proportion of total specimens where the new test and the imperfect standard agree. You can calculate estimated overall percent agreement from Table 3 in the following way.

$$\textbf{overall percent agreement} = 100\% \times (a+d)/(a+b+c+d)$$

Since agreement on absence of disease does not provide direct information about agreement on presence of disease, it may be useful to report two additional measures of agreement.

$$\textbf{agreement of new test with imperfect standard-positive} = 100\% \times a/(a+c)$$
$$\textbf{agreement of new test with imperfect standard-negative} = 100\% \times d/(b+d)$$

As an example, consider the same 220 individuals as before. After all 220 are tested with both the new test and the imperfect standard we have the following results.

Table 4. Example of results comparing a new test to an imperfect standard for 220 patients.

		Imperfect Standard		
		+	–	Total
New	+	40	5	45
Test	–	4	171	175
Total		44	176	220

From Table 4, calculate the agreement measures as follows.

overall percent agreement = $100\% \times (40+171)/220 = 100\% \times 211/220 = 95.9\%$
agreement of new test with imperfect standard-positive = $100\% \times 40/44 = 90.9\%$
agreement of new test with imperfect standard-negative = $100\% \times 171/176 = 97.2\%$

An exact 95% confidence interval for overall percent agreement is (92.4%, 98.1%). Confidence intervals for agreement of new test with imperfect standard-positive and agreement of new test with imperfect standard-negative are not easily formulated because the imperfect standard results are subject to variability and the nature of the variabilty depends on unknown factors.

From Table 4, note that the imperfect standard did not correctly classify all 220 patients. The imperfect standard classified 44 patients as positive and 176 as negative. From Table 2, in truth, 51 patients are diseased and 169 are non-diseased. Since the imperfect standard is wrong sometimes,

you cannot calculate unbiased estimates of sensitivity and specificity from Table 4; however, you can calculate agreement.

There are two major disadvantages with any agreement measure. One disadvantage is that 'agreement' does not mean 'correct'. The other is that agreement changes depending on disease prevalence. We now explore these disadvantages.

When two tests agree, one cannot assume they are also correct. In order to demonstrate this, we need a 3-way comparison between the new test result, the imperfect standard result, and the true diagnosis. A useful way to present the 3-way comparison is like Table 5A.

Table 5A. A 3-way presentation of results comparing the new test, the imperfect standard and true diagnosis.

New Test	Imperfect Standard	Total Patients	True Diagnosis	
			+	–
+	+	40	39	1
+	–	5	5	0
–	+	4	1	3
–	–	171	6	165
Total		220	51	169

From the first and fourth rows of Table 5A, the new test and the imperfect standard agree for 40+171=211 patients, but they agree and are both wrong for 6+1=7 patients.

The other disadvantage with agreement measures is that they depend on disease prevalence. Usually, the agreement between two methods is different in diseased patients versus non-diseased patients. As a result, the agreement between the same two tests can change (possibly a lot) just by changing the proportion of diseased and non-diseased patients in the study patient population. Therefore, it is impossible to generalize the agreement computed for one group of patients to another set of patients unless the disease prevalence is the same.

In order to demonstrate this phenomenon, start with the data from Table 5A. The disease prevalence in this study population is 23.2% (51/220). In diseased patients (Truth+ column), the percent agreement between the new test and the imperfect standard is 88.2% ((39+6)/51), and in non-diseased patients (Truth– column) it is 98.2% ((1+165)/169. The overall percent agreement combining diseased and non-diseased patients is 95.9% ((39+6+1+165)/220), which is the same number computed from Table 4.

To show how disease prevalence affects agreement, suppose that the disease prevalence in the study population is much lower, but the agreement between the new test and imperfect standard in

both the diseased and non-diseased patients remains the same. For example, suppose the study population included 676 non-diseased patients (four times 169) instead of 169 patients so that the disease prevalence in the study population is 7% (51/(51+676)) rather than 23.2%. The new data would look like Table 5B. The Truth+ column in Table 5B is the same as Table 5A, but the Truth– column in Table 5B is four times the results in Table 5A.

Table 5B . A 3-way presentation of results comparing the new test, the imperfect standard and true diagnosis. Disease prevalence is four times greater than that in Table 5A.

New Test	Imperfect Standard	Total Patients	True Diagnosis	
			+	–
+	+	43	39	4
+	–	5	5	0
–	+	13	1	12
–	–	666	6	660
Total		727	51	676

From Table 5B, the percent agreement between the new test and the imperfect standard is still 88.2% ((39+6)/51), and in non-diseased patients (Truth– column) it is still 98.2% ((4+660)/676. However, the overall percent agreement combining diseased and non-diseased patients is 97.5% ((39+6+4+660)/727), higher than the original 95.9%. Showing a more dramatic difference, agreement of new test with imperfect standard-positive is much lower at 76.8% (43/(43+13)) versus 90.9%, and agreement of new test with imperfect standard-negative is slightly higher at 99.2% (666/(666+5)) versus 97.2%.

The performance of the new test and the imperfect standard did not change from Table 5A to 5B, but all of the agreement measures changed simply because the disease prevalence changed. Therefore, it is impossible to generalize agreement measures from Table 4 to another patient population unless you have additional information about disease status (such as Table 5A).

An Example of Discrepant Resolution and Its Associated Problems

As noted before, when a new test is compared to an imperfect standard, the usual calculations from the 2×2 table, a/(a+c) and d/(b+d), respectively, are biased estimates of sensitivity and specificity. Discrepant resolution, described next, is used as an attempt to solve the bias problem. In fact, discrepant resolution does not solve the bias problem; it is just a more complicated wrong solution.

Discrepant resolution is multi-stage testing involving, at a minimum, a new test, an imperfect standard and a "resolver" test (perfect standard). The decision to use the resolver test depends, in part, on the outcome of the new test.

In the simplest situation, discrepant resolution can be described as a two stage testing process in the following manner. In stage 1, test all specimens using the new test and the imperfect standard. The results are presented as in Table 4. In stage 2, when the new test and imperfect standard disagree, run an additional test (resolver) to see which one is "right." Table 6 indicates the retested specimens. The outcome of the resolver is reported in Table 7.

Table 6. Two stage testing process of discrepant resolution. The (discrepant) specimens on the off diagonal (in bold) are additionally tested by a resolver.

		Imperfect Standard +	Imperfect Standard –	
New	+	40	**5**	← Retest
Test	–	**4**	171	
		↑		
		Retest		

Table 7. Resolver results.

New Test	Imperfect Standard	Total Patients	Truth (Resolver) +	Truth (Resolver) –
+	+	40	N/A	N/A
+	–	5	5	0
–	+	4	1	3
–	–	171	N/A	N/A
Total		220	N/A	N/A

N/A = not available

The results in Table 7 indicate that the new test agrees with the resolver (8 specimens) more than the imperfect standard agrees with the resolver (1 specimen) for the study population. However, it is impossible to estimate the relative magnitude of this difference or generalize this difference to a different patient population unless we know the true disease state for all specimens (such as Table 5) or the disease prevalence in the study population.

From a statistical perspective, retesting discrepant results is not necessary. If you do retest these specimens, a good way to report these results is like Table 7. However, it is not appropriate to use the resolver results to revise (change) the original 2×2 table of results because the revision is based on assumptions that aren't verified and usually aren't correct. As a result, it is inappropriate to make sensitivity and specificity type calculations or agreement calculations using the revised table.

Specifically, it has been the practice of some to revise the original 2×2 table of results (Table 4) based on discrepant resolution (results in Table 7). The original 2×2 table is modified using the following (unsupported) reasoning.

- When the original results (new test and imperfect standard) agree, assume (often incorrectly) that they are both correct and do not make any changes to the table.
- When the original results disagree and the imperfect standard disagrees with the resolver, change the imperfect standard result to the resolver result.

Table 8 is an example of how the results from Table 7 are inappropriately used to compute revised results. Specifically, all 40 New Test+/Imperfect Standard+ specimens are incorrectly counted as Truth+, and all 171 New Test–/Imperfect Standard– specimens are incorrectly counted as Truth–. Next, the 5 New Test+/Imperfect Standard–/Truth+ specimens are moved to the New Test+/Imperfect Standard+ total, and the 3 New Test–/Imperfect Standard+/Truth– specimens are moved to the New Test–/Imperfect Standard– total. The 1 New Test–/Imperfect Standard+/Truth+ specimen stays in the New Test–/Imperfect Standard+ total.

Table 8. Inappropriate revision of original results (Table 4) based on discrepant resolution results (Table 7).

New Test	Imperfect Standard	Total Patients	True Diagnosis		*Revised Totals*
			+	–	
+	+	40	*40**		*45*
+	–	5	↑**5**	0	*0*
–	+	4	1	**3**↓	*1*
–	–	171		*171**	*174*
Total		220			*220*

* All specimen results incorrectly assumed to be correct

Typically, the revised totals from Table 8 are presented in another 2×2 table such as Table 9B.

Table 9. Inappropriate revised results (Table 9B) based on discrepant resolution of the original results (Table 9A).

9A. ORIGINAL RESULTS

		Imperfect Standard		
		+		–
New	+	40	← (5)	5
Test	–	4	(3) →	171
Total		44		176

9B. REVISED RESULTS

		Imperfect Standard or Resolver?	
		"+"	"–"
New	+	45	0
Test	–	1	174
Total		46	174

percent agreement = 95.9% (211/220) ≤ apparent percent agreement = 99.5% (219/220)

There are several consequences of revising the original 2×2 table using resolver results. Three consequences are listed below.

1. The columns of the revised table aren't clearly defined and don't necessarily represent truth, as assumed.
2. Calculations of sensitivity and specificity from the revised table are not correct.
3. The "apparent" percent agreement calculated from the revised table will always be greater than or equal to percent agreement calculated from the original 2×2 table.

The third consequence needs further explanation. The agreement calculated from the revised results is called "apparent" because agreement with "what" isn't clear. For some specimens it is agreement with the imperfect standard, and for others it is agreement with the true diagnosis. The reason apparent agreement can only get better is that results can move from the off-diagonal (disagreement) cells to diagonal (agreement) cells in the table, but they can't move from agreement to disagreement. In fact, using a coin flip as the resolver will also improve apparent agreement. Finally, revising results based on discrepant resolution involves using the outcome of the new unproven test as part of the comparative process used to determine the new test performance. This last consequence seems to contradict good science.

In summary, it is not appropriate to revise the original 2×2 table of results based on discrepant resolution because the revision is based on assumptions that aren't verified and usually aren't correct. As a result, it is inappropriate to make sensitivity and specificity type calculations or agreement calculations using the revised table. Instead, FDA recommends reporting the original 2× 2 table of results (Table 4), a description of the imperfect standard, an agreement measure and its confidence interval.

References

Alonzo, Todd A. and Pepe, Margaret S. Using a combination of reference tests to assess the accuracy of a new diagnostic test. *Statistics in Medicine* 1999;18:2987-3003.

Begg, C.G. Biases in the assessment of diagnostic tests. *Statistics in Medicine* 1987;6:411-423.

Bross, I. Misclassification in 2×2 tables. *Biometrics* 1954;10:478-86.

Fleiss, Joseph L. *Statistical Methods for Rates and Proportions*, second edition. New York: John Wiley & Sons, 1981.

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

Green, T.A., Black, C.M., and Johnson, R.E. Evaluation of bias in diagnostic-test sensitivity and specificity estimates computed by discrepant analysis. *Journal of Clinical Microbiology* 1998;36:375-81.

Hagdu, A. Bias in the evaluation of DNA-amplification tests for detecting *Chlamydia trachomatis*. *Statistics in Medicine* 1997;16:1391-9.

Hagdu, A. Discrepant analysis: a biased and an unscientific method for estimating test sensitivity and specificity. J*ournal of Clinical Epidemiology* 1999;52:1231-1237.

Hawkins, Douglas M, Garrett, James A. and Stephenson, Betty. Some issues in resolution of diagnostic tests using an imperfect gold standard. *Statistics in Medicine* 2001;20:1987-2001.

Hayden, Cheryl L. and Feldstein, Michael L. "Dealing with Discrepancy Analysis Part 1:TheProblem of Bias," *IVD Technology* 2000;Jan/Feb:37-42.

Hayden, Cheryl L. and Feldstein, Michael L. "Dealing with Discrepancy Analysis Part 2: Alternative Analytical Strategies," *IVD Technology* 2000;Mar/Apr:51-57.

Hui, Siu L. and Zhou, Xiao H. Evaluation of diagnostic tests without gold standards. *Statistical Methods in Medical Research* 1998;7:354-370.

Lang, Thomas A. and Secic, Michelle. *How to Report Statistics in Medicine*. Philadelphia: American College of Physicians, 1997.

Lipman, H.B. and Astles, J.R. Quantifying the bias associated with use of discrepant analysis. *Clinical Chemistry* 1998;44;1:108-115.

McAdam, Alexander J. Discrepant analysis: how can we test a test? *Journal of Clinical Microbiology* 2000:38:2027-2029.

Miller, W.C. Bias in discrepant analysis: When two wrongs don't make a right. *Journal of Clinical Epidemiology* 1998;51:219-31.

Miller, W.C. Editorial Response: Can we do better than discrepant analysis for new diagnostic test evaluation? *Clinical Infectious Diseases* 1998;27:1186-93.

NCCLS. *Assessment of the Clincial Accuracy of Laboratory Tests Using Receiver Operating Characteristic (ROC) Plots.* NCCLS document GP10. NCCLS, 940 West Valley Road, Suite 1400, Wayne, PA 19087-1898.

Saah, Alfred J. and Hoover, Donald R. "Sensitivity" and "specificity" reconsidered: the meaning of these terms in analytical and diagnostic settings. *Annals of Internal Medicine* 1997;126:91-94.

Schatzkin, A., Conner, R.J., Taylor, P.R., and Bunnag, B. Comparing new and old screening tests when a reference procedure cannot be performed on all screenees. *American Journal of Epidemiology* 1987;125:672-8.

Shoukri, M.M. "Agreement, Measurement of" in *Encyclopedia of Biostatisitics* (Armitage, P. and Colton, T., eds.). New York: John Wiley & Sons, 1998; 103-117.

Staquet, M., Rozencweig, M., Lee, Y.J. and Muggia, F.M. Methodology for the assessment of new dichotomous diagnostic tests. *Journal of Chronic Diseases* 1981;34:599-610.

Tenenbein, A. A double sampling scheme for estimating from binomial data with misclassifications: Sample size determination. *Biometrics* 1971;27:935-44.

Thibodeau, L.A. Evaluating diagnostic tests. *Biometrics* 1981;37:801-804.

Torrance-Rynard, V.L. and Walter, S.D. Effects of dependent errors in the assessment of diagnostic test performance. *Statistics in Medicine* 1997;16:2157-2175.

Vacek, P.M. The effect of conditional dependence on the evaluation of diagnostic tests. *Biometrics* 1985;41:959-968.

Valenstein, Paul. Evaluating diagnostic tests with imperfect standards. *American Journal of Clinical Pathology* 1990;93:252-258.

Walter, S.D. and Irwig, L.M. Estimation of test error rates, disease prevalence and relative risk from misclassified data: a review. *Clinical Epidemiology* 1988:41:923-37.

Zhou, Xiao H. Correcting for verification bias in studies of diagnostic tests. *Statistical Methods in Medical Research* 1998;7:337-53.