JOURNAL OF CLINICAL MICROBIOLOGY, June 2000, p. 2027–2029 Vol. 38, No. 6
0095-1137/00/$04.00+0

GUEST COMMENTARY

Discrepant Analysis: How Can We Test a Test?

ALEXANDER J. McADAM*

Department of Pathology, Brigham and Women's Hospital, and Department of Pathology, Harvard Medical School, Boston, Massachusetts 02115

Nucleic acid testing has arrived in the diagnostic microbiology laboratory, and it has brought along new questions about the statistical evaluation of tests. Few microbiologists are fond of statistics, but we should pay close attention to the use of statistics in the evaluation of new tests: patient care depends on it.

Many of the molecular diagnostic tests used in microbiology include amplification of bacterial or viral nucleic acids. Tests such as PCR and ligase chain reaction depend on amplification of nucleic acid before the detection stage of the test. Nucleic acid amplification (NAA) tests have become common for *Mycobacterium tuberculosis*, *Neisseria gonorrhea*, *Chlamydia trachomatis*, and human immunodeficiency virus (HIV) (6–8, 14). The signal amplification of PCR is extraordinarily efficient, so that even a single organism may be detected, at least in theory. Moreover, because nucleic acid is detected, replication of the bacteria or virus is not needed. Even dead bugs can be detected. These are strong reasons for thinking that NAA tests may be more sensitive than conventional methods, particularly for detection of bacteria or viruses that are difficult to grow. The great sensitivity of NAA tests may increase the risk of false-positive results (15).

The difficulty in evaluating the new tests arises from this quandary: how can a new test, expected to be highly sensitive, be compared to an insensitive, older test? Specifically, what can be done when samples are negative by an insensitive culture method but positive by an NAA test? Many investigators have chosen to perform further testing specifically on this puzzling group of samples; this practice is known as discrepant analysis (4).

Let's take a hypothetical example. Suppose that a new NAA test for disease due to active cytomegalovirus (CMV) is to be evaluated (in this article the "new test" is a test under evaluation, for which the test statistics are being determined). Culture of CMV on cell lines is used as the "gold standard" (the test against which the new test is measured). The results for 1,000 samples tested are given in Fig. 1A. The sensitivity of the new test is equal to the true positives (TP) divided by the sum of the TP and the false negatives (FN) (12), as in the example: sensitivity = TP/(TP + FN) = 155/(155 + 15) = 91.2%. The specificity of the new test is equal to the true negatives (TN) divided by the sum of the TN and the false positives (FP) (12), as in the example, specificity = TN/(TN + FP) = 790/(790 + 40) = 95.2%.

If the specificity of the gold standard test is thought to be excellent (near 100%), the investigators would conclude that the discrepant results in which the NAA test was negative but culture was positive were indeed false negatives for the NAA test. These discrepant results would be accepted, and no further analysis would be done on the samples. While discrepant analysis could include further testing on the culture-positive, NAA-negative samples with a third test, this seems to be uncommon in microbiology (11, 13).

The more problematic discrepant results are the 40 samples in which the NAA test is positive but the gold standard test is negative. If the investigators believe that the NAA test is more sensitive than the old test, they might do an additional test on these discrepant samples. Suppose that a CMV antigen assay is done using these 40 samples and that the antigen test is positive for CMV in 38 of the 40 retested samples. Using the results of antigen assay to create a new "polished" gold standard, the authors would then analyze the data as shown in Fig. 1B. The sensitivity of the NAA test would now be 92.8% (a gain of 1.6%), and the specificity would be 99.7% (a gain of 4.5%).

Is this a reasonable approach? To answer this question, consider what would have happened if a ridiculous test were used to resolve the 40 discrepant results. If a fair coin were tossed to resolve each of the 40 problematic results, 20 of the discrepant results would become "true" positives and 20 would remain "true" negatives. The apparent sensitivity and specificity of the new test would become 92.1 and 97.5%, respectively (improving by 0.9 and 2.3%). In fact, any test used to resolve the 40 discrepant results can only improve or leave unchanged the apparent sensitivity and specificity of the new test (2, 5). Only if no results are reclassified by the resolving test will the sensitivity and specificity appear unchanged. Discrepant analysis, as used in the example, will never reduce the calculated sensitivity and specificity of the new test.

Upon reflection, the reason for this trend is clear. Only results that weaken the sensitivity and specificity of the new test are evaluated by the resolving test. In turn, any changes in the interpretation of the results can only favor the new test.

Even greater changes are made in the calculated positive predictive value (PPV) of the test in this example. Simply put, the PPV of a test is the chance that a patient with a positive test actually has the illness or infection which the test is meant to detect (12). The PPV for the original data in our example is calculated as follows: PPV = TP/(TP + FP) = 155/(155 + 40) = 79.5%. However, after discrepant analysis (as in Fig. 1B), the PPV would become 98.9%, showing an increase of 19.4%.

The PPV of a test is used by clever physicians to decide whether a patient should begin therapy or undergo further testing. If the chance that a patient with a positive test is sick is only 79.5%, the physician might seek further tests or wait and monitor the patient's course, depending upon the clinical situation. However, if a positive test means that there is a 98.9% chance that the patient has the disease in question, it would be rare to seek further testing. The increase in PPV in

* Mailing address: Department of Pathology, LMRC 5th floor, 221 Longwood Ave., Boston, MA 02115. Phone: (617) 278-0317. Fax: (617) 732-5795. E-mail: ajmcadam@bics.bwh.harvard.edu.

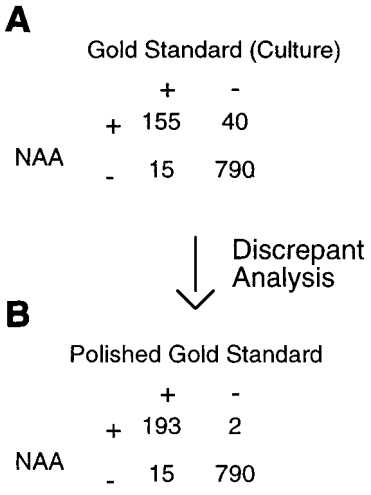

FIG. 1. The effect of discrepant analysis on a hypothetical data set.

the example is smaller than an example from the literature (D. L. McGee and G. H. Reynolds, Letter, Lancet **348:**1307–1308, 1996).

Discrepant analysis will often increase the calculated sensitivity, specificity, and PPV of a test. If performed on the NAA-negative, culture-positive samples, discrepant analysis can also increase the apparent negative predictive value of the new test. Does discrepant analysis makes these figures more accurate than they would be without discrepant analysis? Or is discrepant analysis unreasonably biased in favor of the new test? This issue has caused hot, sometimes almost vitriolic, debate. A number of studies have modeled the effects of discrepant analysis on test statistics, and some trends have become clear.

Discrepant analysis is biased in favor of the new test under most conditions (2, 3, 5, 9, 10). This conclusion is based on studies in which models with estimated test characteristics and disease prevalences are used, and the effect of discrepant analysis is calculated under various conditions. Under most reasonable conditions, discrepant analysis gives higher test statistics (sensitivity, specificity, PPV, and negative predictive value) than the "true" values in the model. In some models the bias of discrepant analysis tends to be small (2), but other models, described below, show that discrepant analysis can cause large biases under some conditions.

The size of the bias caused by discrepant analysis depends on the prevalence of disease (2, 9, 10). At a low disease prevalence, the bias in sensitivity caused by discrepant analysis will be greater. This effect is most pronounced when disease prevalence is below 10% (10), which is common among samples tested in microbiology laboratories. In contrast, the higher the prevalence of the disease, the larger the bias in specificity caused by discrepant analysis. This effect becomes most pronounced when the prevalence of disease is greater than 90% (10), which is uncommon in microbiological testing. In general discrepant analysis is most likely to cause large increases (>5%) in the apparent sensitivity rather than specificity as long as disease prevalence is low.

The magnitude of the bias caused by discrepant analysis also depends on the independence of the resolving test from the new test. "Dependent" tests tend to give the same result, even when the result is wrong (11). For example, two PCR tests, for the same bacteria, that differ only in the choice of primers are likely to be dependent, because contamination with nucleic acid would make both positive, while the presence of a PCR inhibitor would make both tests negative. If the new test and the resolving test are dependent, the bias of discrepant analysis to increase the apparent sensitivity and specificity for the new test is increased (3, 9, 11). The greater the dependence of the new and resolving tests is, the greater the increase in the bias is (10).

It is clear that discrepant analysis isn't perfect. Still, what is an investigator to do when a new test seems to be better than the gold standard? One can certainly sympathize with the desire to accurately portray the value of a new and better test.

I do not have any hard and fast rules to suggest. Instead, investigators (and reviewers) should consider the following suggestions when they find themselves faced with the quandary of a new test that may be better than the old test. It is important to note that the following suggestions are my opinion.

First, pick a gold standard for all the samples and stick with it. If a third test is to be incorporated into the gold standard, use the third test on all the samples. The combination of imperfect tests to form a reasonable gold standard may be the best of bad options (1). Clinical correlation might also be used to determine the true disease state of the patients. If so, get the histories of all the patients. It is true that this will increase the cost and work of performing clinical trials (13; M. J. Chernesky, J. Sellors, and J. Mahony, Letter, Stat. Med. **17:**1064–1066, 1998.), but the money must be weighed against the accuracy of the data. It has been suggested that a random sample of the specimens that are concordant by the new test and the gold standard could be tested by the resolving test, although application of this practice has not been evaluated (10).

Second, think carefully about the choice of tests used in the gold standard. The gold standard should not include tests that are dependent on the new test. If a NAA test is being evaluated, the gold standard should not include a slight variation on the same NAA test. Methods likely to be independent from the NAA should compose the gold standard. Investigators and reviewers will have to use their judgement regarding the independence of tests included in the gold standard from the new test, as the true dependence of tests will rarely be known.

Third, consult a statistician to help design the study. While statisticians can be helpful in evaluating data after a study is completed, they can be more helpful if they are consulted earlier. The definition of the gold standard and the methods used to calculate the qualities of the new test may be improved with help from a statistician.

Fourth, if discrepant analysis is used, the method should be clearly described so that the reviewers can judge if the method is appropriate. The results before and after discrepant analysis should both be provided. Wherever test statistics calculated using discrepant analysis are mentioned, the results obtained before discrepant analysis should also be mentioned. In particular, prominent descriptions of test statistics (for example, in the abstract) should not give only the numbers generated with discrepant analysis. Reviewers can ensure that results calculated using discrepant analysis are presented as reasonably as possible.

These suggestions will increase the difficulty of evaluating new tests. Still, in my opinion, the bias that is inherent in discrepant analysis makes this statistical method unsatisfactory. If a newer, better test requires newer, harder methods of analysis, we are obliged to make the effort to accurately test the test.

REFERENCES

1. **Alonzo, T. A., and M. S. Pepe.** 1999. Using a combination of reference tests to assess the accuracy of a new diagnostic test. Stat. Med. **18:**2987–3003.
2. **Green, T. A., C. M. Black, and R. E. Johnson.** 1998. Evaluation of bias in diagnostic-test sensitivity and specificity estimates computed by discrepant analysis. J. Clin. Microbiol. **36:**375–381.

3. **Hadgu, A.** 1997. Bias in the evaluation of DNA-amplification tests for detecting Chlamydia trachomatis. Stat. Med. **16:**1391–1399.
4. **Hadgu, A.** 1996. The discrepancy in discrepant analysis. Lancet **348:**592–593.
5. **Hadgu, A.** 1999. Discrepant analysis: a biased and an unscientific method for estimating test sensitivity and specificity. J. Clin. Epidemiol. **52:**1231–1237.
6. **Herold, C. D., R. L. Fitzgerald, and D. A. Herold.** 1996. Current techniques in mycobacterial detection and speciation. Crit. Rev. Clin. Lab. Sci. **33:**83–138.
7. **Koumans, E. H., R. E. Johnson, J. S. Knapp, and M. E. St. Louis.** 1998. Laboratory testing for Neisseria gonorrhoeae by recently introduced nonculture tests: a performance review with clinical and public health considerations. Clin. Infect. Dis. **27:**1171–1180.
8. **LeBar, W. D.** 1996. Keeping up with new technology: new approaches to diagnosis of Chlamydia infection. Clin. Chem. **42:**809–812.
9. **Lipman, H. B., and J. R. Astles.** 1998. Quantifying the bias associated with use of discrepant analysis. Clin. Chem. **44:**108–115.
10. **Miller, W. C.** 1998. Bias in discrepant analysis: when two wrongs don't make a right. J. Clin. Epidemiol. **51:**219–231.
11. **Miller, W. C.** 1998. Can we do better than discrepant analysis for new diagnostic test evaluation? Clin. Infect. Dis. **27:**1186–1193.
12. **Pincus, M. R.** 1996. Interpreting laboratory results: reference values and decision making, p. 76–77. *In* J. B. Hery (ed.), Clinical diagnosis and management by laboratory methods. W. B. Saunders Company, Philadelphia, Pa.
13. **Schachter, J.** 1998. Two different worlds we live in. Clin. Infect. Dis. **27:** 1181–1185.
14. **Tang, Y. W., G. W. Procop, and D. H. Persing.** 1997. Molecular diagnostics of infectious diseases. Clin. Chem. **43:**2021–2038.
15. **Vaneechoutte, M., and J. Van Eldere.** 1997. The possibilities and limitations of nucleic acid amplification technology in diagnostic microbiology. J. Med. Microbiol. **46:**188–194.

The views expressed in this Commentary do not necessarily reflect the views of the journal or ASM.