University of Califor
San Francisco

UCSF

Department of Laboratory Medicine
Chlamydia Research Laboratory

5205 '03 SEP 16 P1:34

San Francisco General Hospital
1001 Potrero Avenue
Bldg. 30, Room 416
San Francisco, CA 94110
tel: 415/824-5115
fax: 415/821-8945

10 September 2003

Docket No. 1428
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

Dear Sir/Madam:

I'd like to respond to the FDA document "Statistical Guidance on Reporting Results from Studies Evaluating Diagnostic Tests..." Most of my comments will specifically refer to issues related to diagnosis of *Chlamydia trachomatis* genital tract infections. While some of these comments are likely to be more generalizable, many comments rely on specific lessons learned from evaluating chlamydial tests, and specifically in the context of evaluating a new more sensitive technology in the absence of a gold standard.

We have no perfect test for diagnosing chlamydial infections.

We have no gold standard to use in evaluating new diagnostic tests.

Use of the term sensitivity reflects a moving target. Sensitivity calculations depend on the denominator as provided by a gold standard or composite comparisons. Over years the numbers of positive tests obtained in any specific population with a given prevalence has changed as a function of the performance of the test being used. Thus going from cytologic diagnosis (identifying chlamydial inclusions in Giemsa stained scrapings) to isolation in yolk sac, and then to isolation in cell culture, provided important steps in detecting more infections, and rendered previous sensitivity calculations moot. The introduction of commercial non-culture tests {antigen detection and then nucleic acid hybridization assays} gave us tests that detect dead organisms. and provided a further increment in the total number of infections detected (by culture + the nonculture test) as some specimens from infected

2003D-0044 C7

individuals do not contain viable organisms (usually due too loss of viability in transport???). But good cell culture systems were the most sensitive assays available, and the increment in the denominator (total positives used for sensitivity calculations) was small. Nucleic acid amplification tests (NAATs) obviously provided a much greater increment than the smaller changes seen with most of the earlier technologic improvements.

We have no perfect standard to evaluate new tests. The best evaluations have some degree of compromise.

Discrepant analysis was introduced to provide correct classification of this large number of new positive results being detected (were they TP or FP?)

The FDA document includes assumptions made by those using discrepant analysis that in fact are not correct. These are not the assumptions that were used, nor are the algorithms accurately portrayed. (Parenthetically, was any one doing the evaluations being criticized asked about these?) The basic assumption of the original discrepant analysis was that the specificity of isolation in cell culture approached 100 percent while sensitivity of culture (although the best test available at the time) was lacking. The latter point had been documented in earlier evaluations of non-culture tests. The former assumption is not totally accurate but is pretty close. Clearly there are some false positive cultures, but they are due to cross contamination or clerical errors. The requirement for the original positive NAAT result to be considered a true positive was that it had to be confirmed by another test.

It is clear there is some bias in discrepant analysis and this was clear to everyone doing it. The larger bias is with sensitivity because of the possibility of detecting some positives by retesting previously negative specimens. [That is where all newly introduced, improved sensitivity tests find their pool of positives.... in that group that was previously negative, or the D cell] This is being seen today in a small way when an improvement has been made in a specific NAAT that adds a couple of percentage points in sensitivity, and these positives can not be confirmed readily with the use of single comparison tests. But in the original NAAT evaluations we were faced with roughly a 30% increase in positive results. The bias in sensitivity is not large, as it is based on a small misclassification error and an assumption about undetected positives in the D cell that might be detectable (very technology dependent).

The bias in discrepant analysis calculation of specificity however, is another issue. That bias is very small. The algorithms used for early NAAT evaluations said that a positive result in cell culture was considered as truly positive, and that all other positive results had to be confirmed either by a positive cell culture result or by a second (resolver) test, when the culture was negative, to be considered a TP. This obviated the necessity for testing all negative specimens because a single positive result using a confirmatory resolver could not be considered truly positive (even if it were found to be so, by doing even more confirmatory testing on that specimen, the bias would impact on sensitivity (ie the positive was missed in the first run) not on specificity as the first result was negative). Certainly it is possible that some specimens that are both culture positive and NAAT positive are false positive, but that number must be very, very small and the same is true for the number of specimens that will be positive by two different (either different targets, or procedures) NAATs but are actually FPs. Indeed if these were the

problems we had, and with this degree of error, there would be no need for this letter responding to your guidance document.

The "resolver" test was never considered a perfect test. In fact it was considered to be inferior to the NAAT being evaluated (based on analytic sensitivity). If the resolver had been the better test, the manufacturer would have developed that assay rather than the one that was being evaluated.

Reliance on statistical methods to evaluate diagnostic tests will not provide an answer when there is a dramatic change in the analytic sensitivity of a new diagnostic tests. This is clearly shown by the introduction of NAATs for chlamydia where a threshold for positive signal in the NAATs was between 1 and 10 organisms and the threshold for the previous existing tests was on the order of 10^4 or greater. Given that 20 to 30 percent of clinical specimens seem to fall into the gap between 10^1 and 10^4 there is clear biological reason for an increased number of positives with the NAATs as compared to the other tests. Doing all the statistical evaluations (latent class, or any of the other suggested methods) in the world will provide nothing but an attribution of false positivity to those results when you only have a single NAAT to evaluate, and compare it to less sensitive tests (no matter how many you use). It was necessary to focus on the correct classification of the "excess" positives that led to the use of discrepant analysis

Having four NAATs to evaluate, as we have had recently, clearly reduces any need for discrepant analysis in evaluations, but that is a luxury the original evaluations of NAATs didn't have, only one basic test was available to them. In the future, introduction of any new more sensitive test will raise that same issue again. Are the new positives being detected true or false? In fact think about this: What would happen if a perfect test was actually developed? How would it be evaluated? What performance profile would be calculated? It would have to appear to have a specificity problem, unless one or even two other perfect tests were concurrently made available.

The refusal to accept results from discrepant analysis because of the inherent bias means that specimens that are positive by an NAAT and confirmed by other tests, such as different NAATs, or other target NAATs, or even antigen detection methods, etc, are not going to be considered as true positives except in the very expensive situation where all specimens are tested by all available tests. This leads to package inserts that contain specificity figures that grossly underestimate the true specificity of the test. This, in turn, leads to a situation where any reasonable individual looking at the results presented in the package insert would conclude that NAATs are not suited for screening populations for chlamydial infection. This is the exact opposite of the truth. These tests are very well suited for such use. The basic problem here is a refusal to accept a small bias in specificity calculations inherent in discrepant analysis, but a willingness to accept a much larger misclassification error when results of discrepant analysis are not used.

When a singular new and improved technology is introduced, as when the first NAAT or even second or third were developed, and all such assays were proprietary, cross testing could not be done. There was no option but to use discrepant analysis to confirm the validity of the positives. Discrepant analysis is basically evaluation of excess positives by confirming or retesting.

Witness the current situation with confirmatory testing being recommended by CDC for clinical screening but being rejected by the FDA for test evaluation.

The current situation, combining CDC guidelines for screening for genital chlamydial infection with NAATs together with the package inserts approved by FDA for these NAATs is farcical. CDC recommends using NAATs, and confirming positive results if PPV is <90%. However the specificity figures in the package inserts are so low that achieving such PPVs in the populations being screened is unrealistic. This adds costs, and threatens the viability of screening programs. If one points this out to workers at the CDC, one is told that the "street" knows that the specificity of NAATs is much better than the numbers that the FDA allows. This is bad business, for both CDC and the FDA.

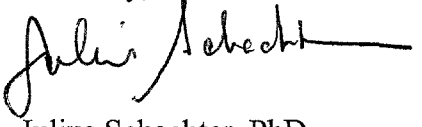
Other comments:

dropping use of terms sensitivity and specificity.....this is attractive, given the fact that true sensitivity is so difficult to measure. But I wonder how the consumers of the tests will respond. So much effort has been expended in educating the users on the meanings of the terms, and their use in further calculations (PPV, NPV) that will no longer be so easy to do.

using agreement as a basic measure.....when I taught basic statistics and epidemiology to medical students, I used to caution them about the use of overall agreement when they read the literature. I used to point out that in the populations where STD screening was usually done (prevalence 3-5%), that I could stand at the door and say "Nope, you ain't got it" and be in 95-97% agreement with the best lab test. The FDA proposal is an improvement as it calls for adding agreement with positive results and negative results. But I think this doesn't go far enough. I would want to know why the disagreement. If it is in the C cell, there is no problem....the new test lacks sensitivity (for lack of a better term). But if the source of disagreement is the B cell. I would want to know if these are TP or FP.....and I would do discrepant analysis to find out. Here my conclusion would be that the new test detects x% more positives than the standard and either: "good news, most are TP, and thus we are detecting more infected people"; or "bad news, these are mostly FP, and.....".

I appreciate the opportunity to comment. I'm sure these remarks will change nothing. But at least now, some of the thinking of those involved in doing discrepant analysis will be in the FDA's hands.

Sincerely,



Julius Schachter, PhD
Professor of Laboratory Medicine

JS/yr