

sgg

ATDEPARTMENT OF HEALTH AND HUMAN SERVICES
PUBLIC HEALTH SERVICE
FOOD AND DRUG ADMINISTRATION

**MICROBIOLOGY DEVICES PANEL
MEDICAL ADVISORY COMMITTEE MEETING**

Wednesday, February 11, 1998

11:00 a.m.

MILLER REPORTING COMPANY, INC.
507 C Street, N.E.
Washington, D.C. 20002
(202) 546-6666

sgg

Room 020B
9200 Corporate Boulevard
Rockville, Maryland

MILLER REPORTING COMPANY, INC.
507 C Street, N.E.
Washington, D.C. 20002
(202) 546-6666

PARTICIPANTS

Microbiology Devices Panel:

Lauri D. Thrupp, M.D., Chairperson
Patricia Charache, M.D.
Paul H. Edelstein, M.D.
Margaret R. Hammerschlag, M.D.

Clinical Chemistry/Clinical Toxicology Panel:

Henry C. Nipper, Ph.D., Chairperson
Arthur Karmen, M.D.
Martin H. Kroll, M.D.

Hematology and Pathology Panel:

Timothy J. O'Leary, M.D., Chairperson
Simon Ogamdi, Ph.D., Dr. P.H.
Margaret S. Pepe, Ph.D.

Immunology Panel:

Glen L. Hortin, M.D., Ph.D.
Henry A. Homburger, M.D.
Sheila Taube, Ph.D.
Mary B. Todd, D.O.
David W. Gates, Ph.D., Industry Representative
Luis A. Rodriguez, M.S., Consumer Representative

FDA Staff:

Steven Gutman, M.D., M.B.A., Director,
Division of Clinical Laboratory Devices
Freddie Poole, Executive Secretary

C O N T E N T S

Opening Remarks and Introductions	4
Conflict of Interest Statement	6
FDA Presentations:	
Opening Statements:	
Susan Alpert, Ph.D., M.D.	7
Gregory Campbell, Ph.D.	11
Overview of Issues, Sharon L. Hansen, Ph.D.	13
Statistical Overview, Kristen Meier, Ph.D.	16
Example #1, Ginette Michaud, M.D.	27
Example #2, Agustin Gonzales, M.D.	31
Example #3, Roxanne Shively, M.T., M.S.	32
Committee Discussion of Presentations	40
FDA Questions to the Panel	57
Open Public Hearing:	
Industry:	
Ms. Linda Ivor, GenProbe	62
Roger Briden, Ph.D., BioStar	66
Professional Group Responses:	
Letters from Julius Schachter, Ph.D. and Wayne C. Miller, M.D., Ph.D., read by Freddie Poole	83
CDC Presentations:	
Timothy Green, Ph.D.	86
Alula Hadgu, Ph.D.	93

P R O C E E D I N G S

Introductory Remarks

DR. THRUPP: I think it is time to call the meeting to order. We apologize for the excess of people and not enough seats, and there may be not enough agendas. I think they are trying to get some more ready for you. If any of you are in the wrong meeting, this meeting was not orchestrated by Kenneth Starr --

[Laughter]

-- despite the fact that if you look on page two of your agenda, Dr. Hagdu is listed as coming from the CDC, Division of Sexually Transmitted Devices --

[Laughter]

-- but a correction to that typo has been passed out.

To lead off, let me just ask the members at the table to introduce themselves, and then we will pass it to Freddie Poole for the conflict of interest review.

I am Lauri Thrupp. I am from the University of California, Irvine, Infectious Diseases Division and Chief of Infection Control at the hospital.

DR. NIPPER: Hello. I am Henry Nipper. I am currently Assistant Dean for Admissions in Creighton University Medical School, and Associate Professor of

Pathology.

DR. O'LEARY: I am Tim O'Leary. I am Chairman of Pathology at the Armed Forces Institute of Pathology.

DR. KARMEN: Arthur Karmen, Albert Einstein College of Medicine. I pledge \$50.

[Laughter]

DR. HORTIN: Glen Hortin. I am Acting Chief of Clinical Chemistry at NIH.

DR. GUTMAN: I am Steve Gutman. I am the Director of the Division of Clinical Laboratory Devices.

DR. GATES: I am David Gates. I am Director of Quality Management and Regulatory Affairs at Becton Dickinson, and I am the industrial representative.

MR. RODRIGUEZ: Good morning. My name is Luis Rodriguez, and I am Assistant Professor of Microbiology at San Antonio College, in San Antonio, Texas, and I am the consumer representative.

DR. EDELSTEIN: I am Paul. Edelstein. I am Director of the Clinical Microbiology Laboratory at the University of Pennsylvania Medical Center.

DR. PEPE: I am Margaret Pepe. I am Professor of Biostatistics at the Fred Hutchinson Cancer Research Center.

DR. OGAMDI: I am Simon Ogamdi, Professor and Chair, Health Sciences, Florida Atlantic University,

Florida.

DR. CHARACHE: I am Patricia Charache. My current title is Director of Performance Improvement Programs, Department of Pathology, where I am Professor of Pathology, Medicine and Oncology, at Johns Hopkins.

DR. KROLL: I am Martin Kroll, also at Johns Hopkins. I am Associate Director of Clinical Chemistry.

DR. THRUPP: I will turn the meeting over for the moment to Freddie Poole.

Conflict of Interest Statement

MS. POOLE: Thank you. For today's joint meeting of Microbiology Devices Panel, the Clinical Chemistry and Clinical Toxicology Devices Panel, the Hematology and Pathology Devices and the Immunology Devices Panel of the Medical Advisory Committee, we have assembled panel members from each of those panels, and we have panel chairs from chemistry and hematology also, sitting on my right.

For today's meeting, the agency reviewed the submitted agenda and determined that no conflict of interest issues are associated with this meeting. In the event that the discussions involve any issues not already on the agenda for which an FDA participant has a financial interest, the participant should excuse himself or herself from such involvement, and the exclusion will be noted for the record.

With respect to all other participants, we ask in the interest of fairness that all persons making statements or presentations disclose any current or previous financial involvement with any firm whose products they may wish to comment upon.

DR. THRUPP: We would like to turn the meeting over to Dr. Susan Alpert, who is Director of the Office of Device Evaluation. Susan?

FDA Presentations

Opening Statements

DR. ALPERT: Thank you, Dr. Thrupp. Ladies and gentlemen of the panel, ladies and gentlemen in the room, let me thank you for attending this very different, if you will, advisory panel meeting. As you are probably already aware, we have combined members from a number of our panels because we have a problem and issue of concern that crosses all of our in vitro diagnostic areas, and we are looking for some help.

We are challenged on a regular basis with new technologies coming in our door that are different, hopefully more sensitive, more specific, of greater value, but not necessarily so. Sometimes simply different: a little more sensitivity, a little less specificity, or the opposite. But new technologies, new approaches to providing

diagnostic information to laboratorians and to healthcare practitioners are constantly being developed, and the way in which we evaluate in vitro diagnostics asks us, in large part, to do some comparisons.

We look at most of these products, or many of these products in the 510(k) arena where we are looking for substantial equivalence. That is a really hard question when you are dealing with totally different technological mechanisms to diagnose a patient, to identify whether a patient has a normal or abnormal result, whether that is from an evaluation of a PAP smear; whether that is whether or not they have a particular infectious disease or are carrying an organism; or whether or not they have a normal or abnormal blood chemistry. If the technologies are different, sometimes we are challenged in looking at the data that comes in where the new technology has identified more patients than what we have considered the gold standard, or identified less patients and narrowed the identified group.

We are challenged to figure out what that means; what is articulated as Truth, with a capital "T", versus, in the environment of comparing the results of two very different tests looking for similar information, how do we make that comparison? Do we make value judgments? We are

frequently asked by companies to allow them to say they are more sensitive and more specific based on simply the laboratory information. Well, is that really true? They may identify more patients but does that make them more sensitive, more specific for identifying patients with or without disease? It is a challenge that we face regularly, particularly as the new technologies are quite different from what we considered gold standards.

What we are asking you, as laboratorians and as practitioners of healthcare, to help us do is to develop an approach that is consistent, something we can use consistently across all in vitro diagnostic areas as an approach to understanding what to do with the information that is provided to us. What does that mean? Are there certain constructs of testing that can help us in providing better information to you about a new test, a new technology? How should we articulate that information in package labeling so that you know how the new test compares to an old and how it compares to other information, or how it relates to other information about the patients that you are involved with, helping to diagnose, helping to take care of, helping to treat.

So, we are asking for your help here. We think this is really challenging. I mentioned this in our senior

sgg

staff meeting at the Center level yesterday, and got a couple of very confused looks and then a discussion of how difficult this is because there is no one answer. We recognize that. I don't want you to think we are asking you to give us one answer on what to do with everything because we recognize that answers will vary depending on the kind of testing, the type of information and the technology.

What we are asking for really is to help us form a sort of global approach. What do we do when we have, for example, a new technology based on genetics, based on genetics technology, that gives us information that is different from--I am a microbiologist; you are going to hear microbiology--from a culture? The gold standard is culture and the genetic-based test identifies 20% more patients. Are they patients we want to manage the same way we manage culture-positive patients? I don't know the answer. How do we evaluate, how do we use the new information and what it means, or how do we communicate that to you, the users, so that the ordering physicians and the laboratorians know what to make of the information?

It is interesting. It is challenging. I think it clearly is of interest to a large group of individuals not on the panel, both from the industry as well as internally, in the agency.

With that, I am going to stop and let you hear from the subject experts within the Division who have provided you with some background and with some case examples of why this is a problem for us. These are not the only issues but we believe that these cases provide examples of the types of problems that we face, the types of questions we are being asked to arbitrate for a manufacturer relative to the other products in the marketplace, and we are looking to you for some help. We are honest. We know there is no one answer but we are looking for you to help us figure out what the global approach would be; what kinds of advice ought we to give to the industry so that the information coming in is useful not only to us but to you because the goal of our review on the information is to provide good diagnostics to you, and good information for your patients. With that, I am going to turn it back to Dr. Thrupp.

DR. THRUPP: Our next speaker is Dr. Greg Campbell. He is the Director of the Division of Biostatistics.

DR. CAMPBELL: Thank you, Dr. Thrupp. I am here representing the Division of Statistics, Biostatistics. We work hand-in-hand with Susan Alpert's Office of Device Evaluation on many of the products that you and the various

panels have seen come before you.

One thing that is important today is that you are going to see a lot of statistical terms bandied about, things like sensitivity, specificity, and prevalence, and predictive value. I hope you don't get lost in the statistical arguments. We value you for your scientific expertise, for your clinical expertise, for your methodological insight. And there are a lot of difficult problems, and Dr. Alpert has mentioned many of them already, where we need your help to figure out how to evaluate tests, to resolve discrepancies in particular. Is there a gold standard? Is there a standard that is close to a gold standard? What do we do in a situation where we have a new technology and there is no gold standard and we think the new technology, the new diagnostic test may be better than things we have used as gold standards in the past? These are very difficult problems and statistics alone cannot help us out of that quagmire.

So, we are looking to you to offer us some insight into exactly how to cope with these problems. As time goes on, we are sure to encounter more of these problems because the technology is ever advancing.

Another question is suppose we have a series of tests but no gold standard, by looking at a number of tests

sgg

can we begin to approximate truth? By looking at a number of tests can we begin to gain an appreciation for what is really going on? These are some of the issues which will come up today, and I just want to second Dr. Alpert's remarks and say that we are glad you are here, and we hope that you will provide us some insight based on your clinical, your scientific and your methodological expertise. Thank you.

DR. THRUPP: Thank you. It might be worth throwing in just one sentence or comment that the issues faced by the FDA, as broadly summarized by Dr. Alpert and Dr. Campbell, are even more heightened under the current era of re-engineering and new guidelines which require that the FDA arrange guidelines before submissions, instead of attempting to retrospectively generalize the kind of problems that we are facing.

Next, Dr. Sharon Hansen, from the Division of Clinical Lab Devices.

Overview of Issues

DR. HANSEN: Thank you, Dr. Thrupp. Welcome, panel members. One or two of you are missing because there was fog in New York, from what I understand, so that planes are late. But they will be here later.

With the November passage of the FDA Modernization

sgg

Act and the incremental implementation of its statutes, the way we have, and the way we will do business in the future will change dramatically. Effective January 21st, 26 Class II devices across the Division became exempt from 510(k) notification.

On February 19th, 130 Class I devices will also become exempt from 510(k) notification within our Division. The remaining 28 Class I devices will be placed in a reserve category, meaning that they will still require 510(k) notification; that they will not be exempt before they can be legally marketed.

This legislative decision, directed at low risk devices, will significantly reduce the time previously committed to review of these devices, and will allow the agency, particularly our Division, to spend more time, hopefully better time, on higher risk devices.

The Modernization Act also requires that the least burdensome studies be required to establish the safety and effectiveness of a new device to support its intended use and indications for use.

Though this advisory panel meeting has been in the planning stages for some time, its occurrence after the passage of the Modernization Act is particularly timely. In our continuing effort to achieve consistency across the

sgg

Division in our review process, we have convened today to ask for your advice, your recommendations, your guidance and suggestions on issues concerning data collection, analysis, and the resolution of discrepant results utilizing sound, scientific and statistical principles to support the indications for use of an in vitro diagnostic device.

The examples that we will present to you, the first two, will be directed, and are intended as a generic sort of view in the sense that they cross all branches within our Division. The third example is a bit more focused, but we expect that issues raised with new molecular technologies and other new technologies that we see coming soon will impact on each branch in our Division.

Each of the examples is representative of part of the information and data submitted to us to support the intended use and indications for use of a new device. Your responses to the questions put before you today will be, I am sure--I am convinced, of considerable help to us as we develop guidance for the IVD industry on the appropriate scientific and statistical methods to incorporate into their study designs and protocols to support the claims they want to make for their product.

The questions at the end of each example will be the focus of what I expect will be a very lively discussion

MILLER REPORTING COMPANY, INC.
507 C Street, N.E.
Washington, D.C. 20002
(202) 546-6666

sgg

this afternoon. In the examples there have been a few changes but the handouts today and the blue-covered copy that the panel has received are a little more up to date. There are not relevant changes but we found typos and we thought of ways to try and present it to you a little more clearly because, as Susan and Greg have both stated, trust me when I say that this is a very complicated issue.

While I have the opportunity, I would like to extend the appreciation of our Division for your continued service on our advisory panels, and we recognize that it is your commitment to the public health, and not the pay, which motivates you.

[Laughter]

DR. THRUPP: Thank you, Sharon. Next we have Kristen Meier, from the Division of Biostatistics.

Statistical Overview

DR. MEIER: Thank you.

[Slide]

This morning I will provide a statistical foundation for evaluating the performance of a diagnostic test. Now, I realize that this presentation is certainly not inclusive of all the complicated issues involved, but I still hope it provides you a framework for discussing what we are all hear to talk about today, which is discrepant

resolution.

[Slide]

So, we begin with the ideal world because if we can't figure it out in the ideal world I think we are going to have trouble figuring it out in the real world. We need to start with a clear concept of who is going to get the diagnostic test. Is it going to be every newborn? Is it going to be those suspected of disease? Is it going to be all sexually active individuals?

This we call the assay target population or specimens from the intended patient population. Once that group is clearly defined, we then have the idea that we want to be able to discriminate between two groups in that target population. We call these groups diseased or non-diseased. Sometimes we call them the condition present group or the condition absent group, or simply the true-positive, true-negative group. Throughout my discussion here I am going to be using the term diseased and non-diseased to be the generic sense distinguishing these two groups.

The clinical question then is to which group does a particular patient belong?

[Slide]

In the absence of a diagnostic test, the best we can do is collect information about the prevalence of the

disease in our target population. Prevalence, which I will just abbreviate as PEV here, is the number of patients in the target population with a disease divided by the total number of patients in the target population, or it is just the chance that a patient in the target population has a disease.

Fortunately for public health reasons, the prevalence is usually low, but we still want to be able to identify those individuals so that we can begin treatment. Ideally then, an informative diagnostic test is one that will identify a subgroup of patients, usually those with a positive test result, in which the disease prevalence is greater than that in the target population.

[Slide]

So, how do we evaluate the performance of a test? Well, first off, we need to evaluate the performance in both the diseased and in the non-diseased group. Again, if it is a good test, the test outcome should depend on whether or not you have the disease. So, it is important to look at both groups.

In the diseased group of patients we typically measure the proportion of test results, and we call that proportion sensitivity. In the non-diseased group we estimate the proportion of negative test results and we call

that specificity--not to be confused with the way these terms are also used in terms of analytic sensitivity-specificity which are limited detection kind of issues. Here, we are talking about the specific definitions, which I give here.

[Slide]

To estimate sensitivity and specificity we collect data and typically present it in the form of a 2X2 table. Here we categorize specimens according to whether they are diseased and non-diseased from the target population. We then categorize the results as either positive or negative by the diagnostic test. From these 4 numbers, here, we can compute an estimated sensitivity and an estimated specificity using these formulas, here.

[Slide]

You will notice that I have in parentheses here that sometimes a positive or negative result is determined through the use of a cut-off value. In many cases the cut-off value is going to give you the plus or minus. If we use a different cut-off value we would get a different 2X2 table and, therefore, a different estimated sensitivity and specificity.

If you imagine all the different possible 2X2 tables you could generate using different cut-offs and

plotting the sensitivity versus 1 minus specificity from all those tables, you would get a curve here. We call that curve a receiver operating characteristic curve, or ROC curve.

This line, this 45 degree angle line, is an uninformative test where sensitivity is equal to 1 minus specificity. The further the ROC curve is for a particular assay, away from that diagonal line, the better discriminator that test is for distinguishing between the diseased and non-diseased group of patients.

[Slide]

Now, in practice we don't know whether an individual is diseased or not. We just have the result from the test. That is why it is also useful to provide additional probabilities that would be useful to a clinician, which would be the positive predictive value and the negative predictive value. I will abbreviate those as PPV and NPV. A positive predictive value is just the probability that a test-positive patient has the disease, and can be computed from the 2X2 table using this formula, here. A negative predictive value is the probability that a test-negative patient will not have the disease. Again, that can be computed from a 2X2 table for the particular prevalence in the study.

The important thing to remember with positive and negative predictive values is that they will change depending on the prevalence of the disease. If the target population has a different prevalence of disease than that used in the study population, then the PPV and NPV computed using these formulas aren't relevant. So, that is why it is also often helpful to actually compute PPV and NPV for a range of prevalences, as you can see, for example, in example 1(c).

[Slide]

Now we finally get to the real world. How do we define diseased and non-diseased in practice? That is, how do we develop the case definitions? There has been a variety of ways that have been used. Some are based on clinical criteria, some on what people have called a gold standard or reference method, or any combination of the above.

[Slide]

I think here is where some of the complications start to come in, or some of the confusion starts to enter. Suppose we have one group that uses a case definition based on an imperfect reference. By reference, I am using it here in a generic sense. It could be a set of clinical and/or analytical criteria. We might have a second group of

experts who might define their case as imperfect reference and a second reference. The question is which is the true sensitivity and specificity? Is it the one from Table 1 or is it the one from Table 2?

Well, in the sense of tracking true disease state, probably neither is the true sensitivity-specificity but both, for practical purposes, could be considered legitimate estimates of sensitivity and specificity. What is critical here is that the case definitions for diagnostic truth be provided; that when we talk about sensitivity and specificity we don't leave off what it is sensitive and specific for. In other words, what are the two groups we are trying to distinguish?

[Slide]

This leads us then into an area that we are here to discuss today, which is discrepant resolution. It may be that one group of experts likes the idea of using a second reference in their case definition but it is simply not feasible or, for whatever reasons, it is not practical to analyze every specimen using this additional method.

What is typically done then in this case is to use some sort of reference for classifying patients as positive or negative and then, on these cases here where the two tests disagree, perform a third test, or in this case a

sgg

resolver, and for now I will assume that it is a perfect resolver when the tests disagree.

What we have then are some of these specimens from the (b) cell, which were negative, which are going to be reclassified as positive based on the result of the resolve, and some of the specimens in the (c) cell, which were positive, will be reclassified as negative based on the outcome of the resolver. These results, here, are considered to be correct. In other words, the imperfect reference is correct when the two tests agree.

[Slide]

The concern with this approach, without additional information, is that the criteria for defining a true positive and negative are not the same for each specimen. Basically, the criteria that you are using depend on the outcome of the new test which really hasn't been proved yet.

If we take as an example the numbers from example 1(b), suppose we had 40 specimens that fell in this category, 5 here, 4 and 171 specimens here, a resolver, or a third test, is used on these specimens, here, and all 5 of these that were negative switch over now to become positive. So, we now have 0 in this cell and 45 in this cell. Three of the 4 that were positive end up being reclassified as a negative, which ends in 1 positive now and 174 in the

sgg

negative category. What happens when you recompute a sensitivity-specificity is that you always get "stay the same" or "improved" when you compute these proportions based on a resolved 2X2 table.

The question I think we need to think about is what is the clinical meaning of these numbers of resolved sensitivity and specificity. What do they estimate?

[Slide]

Sensitivity, with respect to a reference, I think is pretty easy to think about. It is the proportion of patients with a positive new test result out of those with a positive reference result. Resolved sensitivity, on the other hand, is the proportion of patients with a positive reference or a positive new test; negative reference and a positive resolved; out of those patients with a positive reference, or negative reference, positive resolver, positive new test; but not a positive reference, negative resolver, negative new test resolved. Well, if you are having trouble understanding what this means, then you understand the point of this slide, which is that it is not immediately obvious just what this thing is estimating.

[Slide]

So, when we resolve discrepant in this manner with no additional information, we cannot obtain a direct

sgg

estimate of sensitivity or specificity of the new test with respect to either the reference with respect to a resolver or the resolver plus reference, and we can't obtain positive predictive values or negative predictive values.

What we can do from a statistical perspective, however, is to test the null hypothesis, that the sensitivity of the new test is equal to the sensitivity of the reference and, similarly, the specificities are equal.

The catch here is that the sensitivity and specificity we are estimating is with respect to the resolver. The question is, if this is not a perfect resolver, do even these estimates here have meaning? Is that resolver perfect or not? This issue of whether or not the resolver is perfect is really what I see as a separate issue, which the next slide addresses but often gets tangled up in the issue of discrepant resolution.

[Slide]

That is, when we are preparing the performance of a test with respect to another test there may or may not be a perfect method. I will call this other test a predicate method or test. What is done in this case is that you can still use the same formula for sensitivity that we used earlier, but the question now comes in of what is the meaning of that quantity. This term is often called

sgg

relative sensitivity, and it is the proportion of positive new test results in those patients with a positive predicate result. It is really a statement about the agreement of the two assays. No judgment can be made about the assay's ability to predict the condition or disease. So, my question to you is, is that really sensitivity?

[Slide]

Now, there are a lot of issues involved with discrepant resolution and I am certainly not going to address them or even begin to address them, and there is a lot of research going on from members in the audience here and FDA on how to deal with this complicated issue.

I provide this slide just as one example of one of the ways we can proceed but, again, this is not a list or even the beginning of a list of the possibilities and ways we can resolve this issue of discrepant resolution.

One idea that has been used is to randomly sample specimens from every cell, not just look at these cases where the new test disagrees but look at least a sample of all specimens and, from that, one could obtain sensitivity-specificity.

[Slide]

So, in summary, when describing the performance characteristics of a new diagnostic test I think it is

sgg

important that we clearly define the assay target population; that we clearly define the two groups that are being distinguished, in other words, that we provide the case definitions or indicate what the test is sensitive and specific for. Thirdly, I think the study population needs to consist of specimens in both the case-positive and case-negative groups which are sampled from the assay target population. Finally, the clinical and statistical interpretation of the performance characteristic that is reported needs to be clear to the end user. Thank you.

DR. THRUPP: Thank you, Dr. Meier. We are going to have an opportunity after lunch for questioning. However, before we move on, we have a couple more panel members who have become un-fogged or un-rained and have joined us, and I think it would be fair to have them introduce themselves at this point. Dr. Todd?

DR. TODD: Dr. Mary Todd, I am Deputy Director of the Cancer Institute of New Jersey.

DR. THRUPP: And Dr. Taube?

DR. TAUBE: I am Sheila Taube, and I am the Associate Director of the Cancer Diagnosis Program at the National Cancer Institute.

DR. THRUPP: I think those are the only members who have come in since we started. Let's move on to the

first case example, being presented by Dr. Ginette Michaud, from DCLD.

Example #1

DR. MICHAUD: Thank you. Good morning, Mr. Chairman, panel members, members of the audience. The first example to be presented is the case of Class II device, regulated through the 510(k) program.

[Slides]

The device is a qualitative ELISA assay performed on serum samples. While we chose the ELISA technology to construct this example, please bear in mind that a variety of other technologies could have been used to illustrate equally well the statistical principles being discussed today. The issues raised should be considered relevant to a broad range of devices regulated by the Division of Clinical Laboratory Devices.

In this particular example the assay detects an analyte to be used in the diagnosis of a condition or a disease in patients who are suspected of having that illness. I want to present to you three different approaches that have been used by device manufacturers to establish the performance characteristics of such a device.

[Slides]

The first approach is described here under example

1(a). In this scenario the new device is compared to another ELISA assay that measures the same analyte and has the same intended use as the new device.

A study was performed using specimens collected from 23 patients with the disease and 25 healthy donors. It is important to note here that the criteria defining disease status in these patients were not stated by the manufacturer.

The results are presented in this 2X2 table that demonstrates agreement between the predicate device and the new device for 20 positive specimens and 25 negative specimens. The relative sensitivity is, therefore, estimated at 87%, the relative specificity at 100%, and we have percent overall agreement of the new device to the predicate device calculated at 93.8%.

[Slides]

This takes us to the second approach used by device manufacturers to establish the performance parameters of their new products. In this approach, the new device is similarly compared to another ELISA assay that measures the same analyte and has the same intended use. And 220 serum specimens were obtained from the target population, that is, patients suspected of having the disease and for whom the assay would be requested.

The results of both the new device and the predicate assay are illustrated in this 2X2 table. You can see that there is agreement for 40 specimens that tested positive with the predicate device and 100 specimens that tested negative with the predicate device. This allowed the manufacturer to estimate relative sensitivity at 91%. The relative specificity was calculated at 97%, and these results gave percent overall agreement of the new to the predicate device of 95.9%.

In this particular case the manufacturer decided to resolve these 9 discrepant results by using yet another assay that measures the same analyte using a different technology, in this case a commercial immunofluorescence assay.

[Slides]

What we see on the left are the original results. On the right are the results after the third assay has been applied to these 9 discrepant results. What you can see is that in the 5 specimens that tested positive with the new device but negative with the predicate device, all 5 have been confirmed positive by the third assay. Of the 4 specimens that initially tested negative with the new device but positive with the predicate device, 3 of those specimens were confirmed negative by the immunofluorescence assay.

So, this allowed a recalculation of the performance parameters. Relative sensitivity was now estimated at 97.8%, which is an increase from the earlier estimate of 91%. Relative specificity was revised upward from 87% to 100%, and the percent overall agreement between the two devices increased from 95.9% to the present result of 99.5%.

[Slide]

Let's look now at the third approach that is employed by device manufacturers to estimate the performance of their new product. In this case, the new device is not compared to a predicate device. Rather, its performance is compared to the diagnostic gold standard for this particular disease which happens to be the clinical diagnosis.

So, 220 serum specimens were obtained from patients known clinically to either have the disease or to not have the disease. The results of this study are shown here, and allowed the calculation of these performance parameters. We have clinical sensitivity estimated at 94.3% with a 95% confidence interval, as shown here. Clinical specificity is estimated at 93.5% with this confidence interval. Positive predictive value and negative predictive value were 73.3% and 98.9% respectively based on a population prevalence of 15.9%.

[Slide]

The final slide shows how the positive predictive value and the negative predictive value can vary as a function of the disease prevalence in the study population.

This concludes the presentation of the first example. I would now like to turn over the lectern to Dr. Gus Gonzales. Thank you.

Example #2

DR. GONZALES: The submission was a 510(k) for a Class II device. It was related to a marker for an inborn error of metabolism. It was a quantitative fluorometric assay. Whole blood onto filter paper was the specimen of choice, and it was intended to be used for screening for a marker in a neonatal population.

The cut-off value was 3.0 mg, based on 30 years of clinical experience but no well-controlled studies. The prevalence of this disease is 1/80,000 newborns. The intended use was supported by studies in which the new device was compared to a predicate device. A total of 923 consecutive samples was collected from newborns over a period of 3 weeks from blood samples that were spiked with increasing amounts of the marker. All the samples were assayed in both the new device and the predicate device. An equivocal sum was established to reduce the false negatives

sgg

and false positives by prompting the user to rerun the sample. Any initial or repeat sample yielding a result of 3.1 mg is considered as a presumptive positive and must be confirmed in a state laboratory.

Samples with values of 2.1 to 3.0 mg on the initial screen were considered as neither negative nor presumptive positive. These results were in the equivocal sum and were retested with the new device. Values below 2.1 were considered as presumptive negative.

[Slide]

Based upon the results of 923 samples and 2 positive spiked samples, the following interpretive results were obtained for the new and the predicate devices. The predicate device is at the top, and on the left is the new device. There were 19 sample on the equivocal sum, and after retesting 5 became positive and 14 remained negative; 2/3 were confirmed as a positive in the state laboratory.

This finishes my presentation, and the next is going to be presented by Roxanne Shively.

DR. THRUPP: The sound system has been a little bit shy. Could we ask the speakers to speak up into the microphone? Thank you.

Example #3

MS. SHIVELY: I am a poor speaker for speaking up.

sgg

Can you hear me?

[Slide]

Example #3 is a 510(k) submission that represents a device using nucleic acid amplification technology. In this case, we are going to focus on a specific type of device that identifies *Chlamydia trachomatis* nucleic acid in clinical specimens. This particular type of assay would be a Class I.

The intended use for this type of device is for the in vitro qualitative detection of specific nucleic acid sequences from *Chlamydia trachomatis* in endocervical and male urethral swab specimens, and also in female and male urine specimens. This type of assay would be indicated for use to identify *Chlamydia trachomatis* genital tract infections in symptomatic and asymptomatic individuals.

[Slide]

The cut-off for this assay would be established by the manufacturer and based on analytical sensitivity, specificity and cross-reactivity studies done along with preclinical testing of specimens from patients determined to be culture positive or culture negative.

Results for this type of assay would be reported as positive for *Chlamydia trachomatis* nucleic acid if the measurement units were above a given threshold. They would

be reported as presumed negative for Chlamydia trachomatis nucleic acid if the measurement units were below a preset value. Measurement units falling in between these two thresholds would be considered equivocal and would be retested.

[Slide]

This is a complicated table and we are not going to look at it long. I just wanted to use it as an example of the way the data from a clinical study evaluation would be represented in the package insert for such a device.

In this example, swabs, here female endocervical, and urine specimens from both males and females were tested with the nucleic acid amplification assay at up to 5 different test sites.

[Slide]

This is another table showing the patient populations tested broken out by symptomatic and asymptomatic strata. Patient populations that were tested included attendees at STD, OB-GYN, family planning and adolescent clinics.

[Slide]

We are going to leave those huge data sets and try to focus on one element in that data. In order to look at the data categories and simplify the discussion, this table

and following tables will show only the data for female specimens. Sites and patient groups, that is, symptomatic and asymptomatic, may be combined together for discussion only without consideration for poolability.

This table, B-1, shows that part of Table 3 for female specimens broken out by the symptomatic and asymptomatic groups. In the type of study design used in these evaluations, an endocervical or urethral swab--well, in this case for the females an endocervical swab from each patient was cultured. A second swab from that patient was collected at the same time and tested by the nucleic acid amplification assay. For many patients, a urine specimen would also be collected at the same visit and be tested by the amplification assay.

Any positive result by the amplification assay from a patient with a negative culture, in these two columns in the center, would have a DFA test done on the culture transport medium, in the case of the swab specimen, or a residual urine specimen, in the case of a urine sample.

For any of these specimens from these two columns that were negative by DFA the remaining specimen would be retested by the manufacturer using a modified amplification assay that detects an alternate nucleic acid target sequence. Those specimens that are tested by this procedure

sgg

are represented in the last column. In these tables and the following ones, the modified amplification assay is designated ALT.

In data summaries and analyses FDA has considered any positive nucleic acid amplification assay result a true positive if it were verified by culture or by DFA, these two lighter shaded columns. Those nucleic acid amplification assay positive results that were not verified by culture or DFA, this light blue column, and those are the ones that were retested in the alternate target assay, were considered false positives in the summary analyses regardless of the alternate target assay result.

For the data set, you can see that the overall sensitivity for female swab specimens was 29% using this approach; specificity, 98.6%. For urines, 83.3% and 99.0% respectively. Those are the last sensitivity and specificity estimates you are going to see.

[Slide]

Now let's step back and take a look at this representative data site. First, the overall prevalence in the female patient population groups by specimen type, endocervical swabs and urine, both symptomatic and asymptomatic patient groups. Here is one of those typos that was found since you got your copies. Please note that

sgg

for the endocervical swab samples in the asymptomatic patient group the prevalence by culture is 4.7%, or 44/934 specimens.

Overall prevalence for females did vary according to the specimen type tested, as you can see in this table, and also the patient population tested. One immediate observation is that testing with the amplification assay has more positive results than by culturing regardless of the specimen type. For the symptomatics, the endocervical swabs, there are 60 positives by the amplification assay, 36 positives by the culture assay.

Of interest with this data would be to determine if that larger number of positives with the amplification assay represents infection in the populations tested, or whether they are false positives.

[Slide]

Another way of grouping the study data is shown in this table. It shows the nucleic acid amplification assay results grouped by culture status. The top grouping are those specimens that were culture positive with the appropriate amplification assay results.

For the group of specimens that were culture negative, we have a breakout into several subgroups. That group of specimens was positive in the nucleic acid

sgg

amplification assay, tested positive in the DFA test, and then those specimens that were positive in the nucleic acid amplification assay again but tested negative by DFA. This is the group that was subjected to further testing with the alternate target assay. Then we do have this group down here that are culture negatives but also were negative in the amplification assay.

[Slide]

Many patients had both a urine and a swab tested by amplification assay in this study design. Table C is a similar breakout for specimens tested from 1,360 females who had both a swab and a urine specimen tested with the amplification assay. Performance of either specimen type could be assessed based on patient infected status. This approach considers a patient infected if either, and that is the column over here, a urine or endocervical specimen tested positive with the nucleic acid amplification assay and was verified by culture or DFA.

In this approach, some specimens that were considered true negatives in the amplification assay for either of the specimen analyses, either endocervical swab or urine, could then become false negatives because one of the matched specimens was positive in the nucleic acid amplification assay. There are 9 specimens in the

sgg

endocervical analysis and 13 in the urine analysis that fell into this category. These are culture negatives and nucleic acid amplification assay negative in the matched patient specimen.

[Slide]

For the analysis of this study data using this particular study design there are options for presenting the amplification assay performance, projecting conventional sensitivity and specificity estimates.

First, we could compare the amplification assay to culture and recognize those amplification assays with verified positivity in the DFA as being true positives. This is the approach that was used in Tables 3 and 4.

Another approach is expanding the gold standard, and this would include those amplification assays that were positive and also tested positive in the alternate target assay. These would be considered as true positives.

A third approach would be to compare both urine and endocervical swab specimen testing results to patient infected status, using the approach in either number one or number two.

This is a very complex data set. I promise not to show 2X2's for sensitivity and specificity estimates. I didn't show any 2X2's or estimates because I thought it

sgg

would be important to focus on the actual specimen testing groups that came out from the study. If you have any questions on any of these examples, we will be available to answer them, and I will turn it back over to Freddie. Thank you.

DR. THRUPP: Thank you, Roxanne. I believe we have a few minutes before the lunch break and, with Freddie Poole's permission, I believe I would like to ask the presenters to remain close to a microphone or be at the table somewhere close by, and we can see if the panel members have any questions to ask concerning the presentations we have just heard. After lunch we will be reviewing the specific questions that Dr. Hansen will review with us, that FDA is asking, and there will be opportunity for more discussion but, since we have about 15 minutes, may we ask for questions from the panel at this point of the presentations? Dr. Charache?

DR. CHARACHE: Just one question about the last one where an alternative sequence was used for the nucleic acid detection. I would wonder what is known about the alternative test; how we know that that alternative test is specific as well as sensitive. Is there information provided about the alternative test that would permit the assessment?

DR. HANSEN: No, there was no information provided in the submission to verify or validate the alternate method. It was assumed it was the same as the original as far as processing but the target sequence was different. Roxanne, do you have anything to add?

MS. SHIVELY: Just to clarify that point, Dr. Charache, that assay is the same assay as the one that was being tested but with a substitute probe, whatever, primers, to detect an alternate target, and those tests have not been marketed commercially.

DR. THRUPP: Since we are on this last example, could I throw out one more question for Roxanne on the CT procedure? Since we have to focus on target populations, were there included in the material definitions of the target population? For example, were patients who had been on treatment or recent treatment excluded?

MS. SHIVELY: In these types of study designs, yes, there were exclusion and inclusion criteria. Patients who were on antibiotics were generally excluded. Also, the definitions for stratifying patients into symptomatic and asymptomatic groups were defined.

DR. THRUPP: Dr. O'Leary has a question.

DR. O'LEARY: This is a question that is going to come with a statistical example to illustrate the question.

The question is to what extent do you think this problem arises from the binning of continuous data? The example to focus the question is, suppose I put forth the hypothesis that all men are less than 100 ft tall, and I go along and I do a sampling of the population and I find 990 million men who are less than 8 ft tall, and I discover a 99 ft man. This 99 ft man clearly, from a statistical perspective with the question that I asked, tends to support the statistical hypothesis that all men are less than 100 ft tall, but when I inspect the data it really brings into question the problem. It seems to me that seems to be part of the problem here and I am wondering whether you think this question of binning continuous data is part of what we are having to deal with.

MS. SHIVELY: Could I ask a statistician to answer that question?

DR. MEIER: Let me just clarify. I mean, there are two issues, two binnings going on, if you will, or two categorizations. One is your reference or the two categories you are trying to distinguish, and the question is are there really just two categories? The premise, when we even talk about sensitivity and specificity, is that we can conceptualize two. I know that is an issue. It is not always possible to say someone is diseased or not diseased.

There are various stages of disease. That is an issue.

There is also the issue of binning with a single cut-off, and I think that is why the ROC curve provides really more informative information than just a simple 2X2 table because it gives you the entire performance through that.

I haven't even touched the one on the issues of the disease status not really being two categories. Most of the framework, when we talk about sensitivity and specificity, is with that in mind.

DR. O'LEARY: A follow-up though on the ROC curve, because in many cases with a relatively small number of patients that we are illustrating with, you can't really construct that ROC curve with a great degree of confidence. In fact, I haven't seen confidence intervals ever presented on an ROC curve in a medical investigation. The question really comes back again to the statistics. We are really attacking this problem with the statistics of Ronald Fisher and I am wondering whether we shouldn't be attacking it with the statistics of a Bradley Ephram and really be thinking about the problems differently.

DR. MEIER: I am not sure how to answer that. I think people are thinking about these issues. They are not simple issues. I think it is going to come from

sgg

statisticians working more closely with the clinical and research community on that. I think people have addressed this. There are confidence intervals, by the way, for ROC. Unfortunately, software hasn't quite gotten up to speed on doing these but the methodology is there; it is just matching it and making it available.

MS. SHIVELY: Taking your question back to a clinical perspective, or one that is specific to the Chlamydia situation, in those breakouts on specimen results we have a fairly confident feeling about the specimens that are culture positive in that culture is considered, for all practical purposes, 100% specific, or at least up there, whereas, a negative culture result would have a lot less confidence. So, you could look at the groups of specimen test results and have a scaled degree of confidence in the predictability of each of those.

DR. NIPPER: I wanted to add to the confusion being strewn by Dr. O'Leary --

[Laughter]

-- because one of the things that has bothered me a lot about the application of Bayesian statistics and that I would like to throw out for possible consideration by a statistician is that many times test populations are not selected from populations that mimic what the clinician is

seeing in the office. I think that she defined that problem quite well when she said "in people without the disease or from a normal population." I would like you to talk a little bit more about selection of the test population from people with symptoms that may have the disease, just like a clinician would see in the office.

DR. MEIER: Again, I guess when we think about what the physician would use, it would be, say, a predictive value to evaluate given a positive result--what is the chance my patient has disease? That is dependent on both sensitivity and specificity. Well, if specificity, for instance, is measured in the wrong group, in other words normals who wouldn't normally even get this test, then I would just say that your predictive value just doesn't have meaning or it is not necessarily related. I don't know if that is your question.

The point is that you need to select specimens--and I am not saying this is easy, but the idea is that you select specimens as close as possible to the kind that would be in your target population. I think that is the group that you need to estimate the sensitivity and specificity for. That is the group that then you can think about these positive and negative predictive values making sense so that you can interpret the test results and then go

from there.

DR. NIPPER: So, now my question is do you enjoy seeing statistics used in situations--are you professionally rewarded in seeing statistics used in situations where people have just gone to the specimen library and brought out N specimens from box 1 and N+1 specimens from box 2, thrown that in and then were calculating sensitivity and specificity? Is that the way it should be done? Or, how do you feel about it?

DR. MEIER: Again, I guess I would throw that question back to you. If that is the kind of specimens that you are going to typically target for this assay, then yes. But, to me, the question seems pretty straightforward that, no, you need to know something about the specimens.

Margaret, I know you are also a statistician, if you would like to comment on that? The point is the specimens that you are estimating the performance characteristics on have to be typical of that. Again, I understand that is not always an easy thing to do. It is easy for me to say that, but I think an attempt at least has to be made to get specimens from the right population.

DR. THRUPP: Let's go to Dr. Pepe and then we have a question from Dr. Hortin.

DR. PEPE: I agree. These studies have to be

sgg

designed to answer clinically relevant questions, and it is not clear to me from some of these examples that the clinically relevant question is even stated. There is an analogy with the evaluation of new treatments by the FDA. These seem more like Phase II studies and Phase III studies. Phase III studies are supposed to address the clinically relevant impact of treatment, as I understand it.

To get back to your question, Dr. Nipper, about the transportability of sensitivity and specificity from one population, a study population, to another population, I am skeptical about that. There are lots of factors that can affect the sensitivity and specificity of the test, for example, the severity of the disease in the population. In most diseases there is a range of disease severity or load of disease, and we can't really just use the sensitivity from a population that is severely diseased and use that in a population where the disease is not so severe.

So, I think it is important in these studies that we determine what factors are affecting sensitivity and specificity so that we can then determine the real predictive values in populations with known covariates, if you like, where a covariate might be disease severity or other factors that can affect the test.

DR. HORTIN: I just wanted to point out, or maybe

muddy the waters a little bit more, but point out the kind of three different basic problems we are dealing with here that maybe have been touched on or maybe haven't been dealt with.

First of all, kind of the sensitivity and specificity in terms of their clinical use: Oftentimes we think of them in terms of the sensitivity and specificity of the test but really, as we have just been discussing, there are population characteristics, and you can design a study by selecting certain populations to make a test perform very well or very badly depending upon your population selection. So, really oftentimes we think of those as kind of absolute characteristics of the test but really they are very much a population characteristic. So, that is one factor that we have to consider, and it was discussed a little bit in terms of the selection of the population.

The other is that in general the test has not been applied to a population but is being applied to the diagnosis of a specific individual. So, you are taking those population characteristics and they may or may not apply unless the individual is very typical. Those populations that the information was derived from five years ago are not homogeneous and the risks of different individuals in there for false positives, false negatives

and things are not necessarily uniform, but statistically we have generally treated those as homogeneous populations. Most individuals probably are not going to represent the average risk for false positives or for predictive values.

So, the clinician is basically faced with the problem of taking these statistical measures of sensitivity and specificity and applying them to any individual. If he has a 90% sensitivity for picking up something, he is going to decide for his patient yes or no rather than 90% probability or 10% probability of not having the disease.

The third complication I think that has been dealt with maybe a little bit more on the immunology panel is that many tests are not necessarily used purely for diagnostic applications. Many of them are used for monitoring applications, and the application of a population-derived cut-off may not be relevant for monitoring applications. In that setting each individual is basically his own baseline. So, having a requirement for a lot of data about population characteristics for determining a cut-off may be merely irrelevant for monitoring applications and, basically, what you are interested in there is deriving data as to what magnitude or change versus the patient's baseline.

So, those are just kind of three issues that were not kind of discussed very fully and I just wanted to bring

sgg

them up for consideration.

DR. THRUPP: Dr. Hammerschlag has had a chance to catch her breath after her taxi ride. So we would like her to introduce herself for the group.

DR. HAMMERSCHLAG: Margaret Hammerschlag, at the State University of New York Health Science Center, Brooklyn.

DR. THRUPP: We have other questions. Dr. Hansen?

DR. HANSEN: A comment I would like to make, and I think it is relevant to what everyone has said, is that there is an additional statute in the new law that is going to permit the FDA to get together with the companies early on to help design studies.

The study that you have all touched on were some of the issues that we certainly hoped would be discussed today, that is, the study design is probably the most critical part of the initiation of what you are going to do, where we agree on patient populations; we agree on the kinds of individuals who should be included, and that is going to depend on the analyte.

So, consider that if things are done up front correctly, it is a lot easier for everyone. And we certainly, in the Division, are going to be moving in that direction because the new law offers the industry the

opportunity to meet with us before they start their studies so we can agree on the appropriate approaches. That has been done to a limited degree over the years, but we now have a law to permit us to do it. So, these issues that you are bringing up will be very, very helpful to us as well as the industry.

But what we have presented to you today is the way things come in to us and then we are faced with trying to fix it, or upsetting the industry if we need additional studies. If we don't have the studies, what can be said? And those questions will be some that you will be addressing this afternoon. We get what we get, and if it isn't what we think we need, it is very difficult not only for the industry but for us. So, the up front discussions are the ones that I believe will be very, very critical.

DR. EDELSTEIN: I have a question regarding the discrepant pair analysis. In the first two examples given, in fact, there was no significant difference between the discrepant pairs, 0 and 3, and 4 and 5. If the charge of the agency is simply to demonstrate equivalence, then should these insignificant discrepant pair analyses be taken further, or should it just be judged that the tests are equivalent? Are these primarily studies that are done for purposes of marketing and clinical efficacy, or are they

needed to demonstrate equivalency?

DR. GONZALEZ: In this particular case we were dealing with a screening tool for neonatal testing of low prevalence diseases. Perhaps the numbers were sort of made up to make you understand some of the limitations that we are dealing with. The major problem in this kind of situation is that the device touches the lower end of the diagnostic range which, by knowledge, these kind of devices, they have imprecision right there. So, the imprecision is perhaps the one that impacts the most in the interpretation of the results.

So, when we were dealing with the manufacturers and they were coming in with their data, we noticed two things: that moving the target of the cut-off value would change the percent of false positives or false negatives. The main problem that we were dealing with was the imprecisions of the test. They were quite significant, and all these technologies out there, they don't match each other because they use different technologies and they don't match very well with each other. So, we were confronted with problems and limited factors, the imprecision of the test and the issue of moving the cut-off value.

So, we decided that perhaps by creating a grey zone that we called the equivocal zone--we consulted a

statistician, John Dawson and the other person that participated in this idea was Carol Benson, and the three of us, we decided to create this equivocal zone. Our concern was that we didn't want to have babies tested and miss them just because of technicalities in the assay or because of lack of sufficient amount of results from a very low prevalence disease population. So we decided to create this equivocal zone as a safety caution so the user will retest and be sure that the chances of having a false negative would be minimized by the creation of this equivocal zone.

DR. THRUPP: Dr. Gutman has a question, but I would just like to throw out --

DR. GUTMAN: One comment actually. I would like to elaborate on what Dr. Gonzalez has said.

DR. THRUPP: Well, let me add one comment or question that you can both respond to. On this example of the very low prevalence, neonatal 1/80,000 type prevalence, is there information--there was not in the example--about the predicate device and how the decision was arrived at where it does not have an equivocal zone? Then the second aspect of this, the presumptive positive by whatever test is referred to the state laboratory, and how does the state laboratory decide what the gold standard is? Is it merely a test of replicability of the same sample using the same

sgg

predicate device? These issues, it seems to me, are critical in this sequence of decision making.

DR. GUTMAN: I just wanted to point out, before we get too tied up in the particulars, that we were trying to make up examples here to spark interest in the issue of discrepancy and the issue of how to describe things. I, frankly, don't care what the examples are or what the numbers are. You can make them up as you go along or change them to highlight bin issues or other issues that might be of importance to you.

What is much more important to me than any of the examples or resolving any of these specific issues, because they will vary so much from product to product, is to talk in more general terms about what tools we have to reach the challenge that Dr. Edelstein suggested, which is to show equivalency and then to communicate relative performance or superiority, if that is what the company wants to claim, or inferiority, if that is what it appears to be but it is good enough.

And, the real issue that appears, and Dr. Gonzalez was intimately involved in generating the idea, is what the hell do we do when we get a case where there is 1/100,000 or 1/50,000 and we can't possibly do a study on the population and intended use unless we ask the manufacturer to do a

sgg

10-year, 10-site study? Then, do we go back and pick positives from banked samples that are totally biased or partially biased or partially characterized? Do we make up samples using spiked knowns, or what do we do when the reference method is universally used and, in fact, there is no answer to what truth is and it is simply a matter of practice? You won't fall out of your chair if I tell you that that is the case for some methods.

So, what we are here to do is pose general issues and ask for general advice, and the critical issue is how to take a data set, whether it is not enough numbers or too many numbers or great numbers, and what are our choices and what can we offer the manufacturers in terms of meeting the least burdensome threshold but providing information that gives insight to you as laboratorians and clinicians?

DR. THRUPP: We have time for two more questions from Dr. Charache and Dr. O'Leary, and then I believe it will be lunch time.

DR. CHARACHE: To get back to a very generic question then, I am wondering about the rules or criteria that are employed for things like spiked samples. I can see that it may be essential in helping to set up a test initially or working with this 1/80,000 problem. But I would wonder whether the criteria for the use of spiked

specimens might vary, for example, with availability of true samples and perhaps method. I can see that if you are doing mass spec HPLC you are okay with spiked samples in a setting in which metabolites could interfere with the endpoint if you are using an immunodiagnostic technology. I am wondering how all that gets sorted out.

DR. THRUPP: I will take that as a rhetorical question for answering this afternoon. Let's go on with Dr. O'Leary.

DR. O'LEARY: Maybe this is rhetorical as well. We have had an opportunity to read a number of papers out there. This is to the statistical people again, is there anybody out there in the statistical community that has an argument that there is no bias associated with the use of this technique, or a good cheap fix to make it nearly unbiased?

DR. PEPE: The answer to both of those questions is no, as far as I know. There are those of us who have already thrown up our hands and said that this is a problem that can't be solved, like Colin Begg, and I know of no quick fix.

A basic issue when you think about the resolver test is how correlated is that with, say, the new test. I mean, suppose it is making exactly the same mistake as the

sgg

new test is making, then you are gaining nothing except confidence for incorrect conclusions that the new test --

DR. O'LEARY: But even if it is uncorrelated, it means a biased approach --

DR. PEPE: That is right.

DR. O'LEARY: It is inherent bias, regardless of whether or not it is correlated. So it could have absolutely zero correlation and still be biased.

DR. PEPE: That is right. I think there is hope if you can really assume that there is an unbiased mistakes that the resolver can have. In that case, there probably are statistical techniques that can be developed for that problem. It seems like you could identify the sensitivity. A fix-up would be required, of course, because there is bias, as you say, in the way that the methods are employed right now but in real situations you have to think about whether or not that assumption about uncorrelatedness of mistakes is valid.

DR. O'LEARY: In fact, basically to feel comfortable about it you would have to prove absence of correlation rather than fail to prove correlation, and do the opposite of what we normally try to do.

DR. PEPE: That is pretty impossible to do when you don't know what the gold standard is.

sgg

DR. THRUPP: On that fairly clear resolution of all our problems, I think we need some lunch. I understand that lunch is right here. So, we will reconvene at 1:30.

[Whereupon, at 12:30 p.m., the proceedings were recessed to be resumed at 1:30 p.m.]

AFTERNOON SESSION

DR. THRUPP: Let's move on to Dr. Hansen, who will review the questions being posed for the discussion.

FDA Questions to the Panel

DR. HANSEN: The questions have not changed as they were sent to you. So, we will merely review them. But at this point in time, I think I wish Oprah Winfrey were here to perhaps talk about mad cow disease --

[Laughter]

In example #1 there were actually three scenarios, and we are asking your advice on what claims should be allowed for each of the examples, a, b and c. And when we say claims, we are talking about the intended use primarily of the device, how it is written.

How should the performance characteristics be presented in the package insert for each of those scenarios?

How and when should specimen results from "healthy donors or normals" be used?

Should the terms relative sensitivity and specificity be replaced with some other term, such as comparative or estimated sensitivity and specificity, or perhaps a better term for overall agreement from device to device because the three examples within example #1 showed device to device comparison, device to device in an

sgg

appropriate target population, and the third example is one in which we have clinical truth.

However the test result is obtained, should that report to the clinician reflect the nature of the performance characterization? If so, how should it be stated?

As presented, are the resolution of discrepant results statistically and scientifically appropriate?

In the three examples how should and can predictive values be estimated?

It is up to the panel, but the open public hearing will be before your deliberations. So, let's go on to the second example. These are in the handouts from today that everyone in the audience should also have a copy of.

What is the appropriate presentation of the new assay's performance characteristics in the product insert?

For example, number two--and this could apply to any assay that may have an equivocal zone, how should sensitivity, specificity, whether it is relative, clinical or some other term or agreement, be determined for a new device when the new device has an equivocal zone? Would your recommendations change when doing a device to device comparison; a device to a gold standard or reference method; or a device to clinical diagnosis or clinical truth?

sgg

In example 2, how should and can predictive values be estimated?

Based on the discussion of these questions and your recommendations, how should FDA convey this information in the product insert?

The first question in 3 directly applies to Chlamydia devices. To date, we have cleared three different manufacturers' assay using nucleic acid amplification for diagnosis of Chlamydia. For future evaluations of any Chlamydia device specifically, because we know this may vary depending on what the analyte is that the nucleic acid amplification assay is going to detect and other diagnostics where more sensitive technologies are being used, should the gold standard be redefined to include testing by a nucleic acid amplification test that is commercially available or well-described and its performance validated, or should culture continue to be used as the gold standard? In the past the gold standard has been culture for us, with a secondary standard of DFA.

Should we continue to use culture as a gold standard with retesting of culture negative/new assay positives with a different assay, such as DFA or another nucleic acid amplification technology that has been cleared or approved for that particular analyte? Or, should we

sgg

permit nucleic acid amplification of an alternate target site or sequence in the organism using the same nucleic acid amplification technology?

Should we reconsider the gold standard to be detection of specific nucleic acid sequences using a cleared nucleic acid amplification device as the comparative predicate?

How and when should discrepant results be resolved?

The second and third questions deal more with what we perceive to be, and what we know are going to be future submissions, which Susan alluded to this morning, where new technologies are perceived to be better than anything we have in the marketplace now.

We are expecting submissions using nucleic acid amplification technologies, as well as other molecular-based methods, for detection of etiologic agents other than the three that have been cleared or approved by the agency, which include Chlamydia, Neisseria gonorrhoea and Mycobacterium tuberculosis complex. But we perceive that these newer technologies are also going to be used for other intended uses, other than infectious disease diagnosis.

We ask what you might recommend to establish the performance characteristics of such devices when: a) no

sgg

culture methods are available or are impractical even when there are culture methods?

b) When culture or other laboratory methods are not reliable for establishing a definitive diagnosis?

c) How should positive and negative predictive values be estimated for such devices? When would discrepant result resolution be applicable?

Finally, after you find the answers to all these questions, what are your recommendations for presentation of performance characteristics in the product insert under the above scenarios?

Finally, the last question if the moon hasn't come up, when can a device make claims superior to the predicate device, gold standard or reference method when the new device is compared to a predicate? The new device is compared to a reference method or gold standard? Or, the new device is compared to clinical diagnosis or clinical definition using specimens from patients with and without the disease or condition?

We look forward to your answers and suggestions and the resolutions of our discrepant condition. Thank you.

DR. THRUPP: Thank you for those simple questions. At this time, we want to open the public hearing section of this meeting, and we have several individuals who have

sgg

submitted their request and data that they would like to review. The first, which is listed in your agenda, is Linda Ivor, from GenProbe.

Open Public Hearing

Industry

MS. IVOR: My name is Linda Ivor, and I am a clinical scientist at GenProbe, Inc., in San Diego.

[Slide]

GenProbe develops and manufactures genetic probe assays that include amplification tests for Mycobacterium tuberculosis and Chlamydia trachomatis. GenProbe was pleased to learn of FDA's interest in reviewing protocols for data collection, analysis and resolution of discrepant specimen results, and appreciates this opportunity to give comment.

My comments today are directed at the resolution of discrepant specimen results. The purpose of laboratory tests is to aid the clinician in forming an accurate assessment of a patient's health or disease. In a continuing effort to improve the availability and effectiveness of a test, developers have explored and provided new technologies that have the potential of having greater accuracy than older tests. In fact, most new technologies target areas where there is a need for greater

sensitivity.

However, in seeking FDA clearance to market new products, developers must compare their product with an established gold standard that may have less sensitivity. As a result, a new product may appear to have a low specificity as an artifact of the comparison even though it more accurately reflects the patient's health status. For example, culture methods remain the gold standard for Chlamydia tests even though the culture lacks the sensitivity of the newer nucleic acid amplification tests. Because of this disparity, the comparison creates false-positive results that may, in fact, be true positives. To better describe the clinical value of a new test, discrepant specimen analysis is necessary to resolve any comparison bias due to a less effective predicate.

[Slide]

GenProbe wishes to present these items for consideration: First, analysis of discrepant specimens require alternate test or tests of similar sensitivity and specificity. Specimens that are found discrepant relative to the gold standard are considered false positive. Because the new test has the potential of being more sensitive than the gold standard, it is critical that these discrepant specimens are analyzed to ensure they are not true

positives.

In some cases analysis of the false-positive specimens cannot be fully resolved by using other methods that have been cleared by FDA. The need then is for another test that is a more effective comparator assay. For example, GenProbe has received clearance for an amplified test for Chlamydia. The analyses of the clinical data included the use of cultures as the predicate, plus DFA and DFA on matched patient urine for discrepant analysis testing. These tests were effective in resolving some but not all of the initial MCT results.

To address the remaining false-positive specimens, GenProbe introduced an alternate amplification research test. FDA agreed to integrate the use of the amplification test into discrepant testing. Because the alternate test has applied amplification techniques that were more comparable to MCT, it was effective in resolving more of the discrepant results than any of the other methods used in this analysis.

Alternate methods should be allowed for the discrepant resolution for nucleic acid-based tests, and the final sensitivity and specificity of the new test should be calculated with these results.

Second, patient diagnosis, that is, non-laboratory

clinical data should be considered in determining the presence or absence of disease. The product claim should reflect the performance of a test as determined by all that is known about the patient's health status, that is, the gold standard results, patient diagnosis and the discrepant specimen analysis.

[Slide]

For example, GenProbe learned from its IVD test for tuberculosis that culture had insufficient sensitivity and specificity to evaluate a nucleic acid-based test. As a result, and with support from FDA, GenProbe is now pursuing clinical trials using physician-established patient diagnosis as the gold standard. Clinical diagnosis is the medically established standard for ruling TB in or out and, therefore, is a more suitable comparator to establish safety and effectiveness.

In addition, using this clinical standard does not require or allow further discrepant resolutions since it is the medical gold standard.

Third, there is no added benefit in testing all clinical trial specimens with the discrepant specimen algorithm. Mathematically, discrepant testing is biased in the statistical sense because false-positive specimens are selected for additional testing. However, as was presented

sgg

in a recent journal of Clinical Microbiology article by Green et al., the size of the bias appears to be negligible with respect to sensitivity. In fact, the bias decreases as the sensitivity of the new test increases, and the magnitude remains modest provided the true sensitivity is 90% or greater. This is the case for nucleic acid amplification assays.

Finally, this meeting should not be the basis for a single guideline on discrepancy analysis. Because of the singular nature of many technologies, the disease processes and their associated analytes, and the effectiveness or the ineffectiveness of the gold standard as a predicate, the approach to each study design and discrepant resolution should be handled by the developer and the FDA on a case by case basis. Thank you.

DR. THRUPP: Let's hold the questions for Dr. Ivor until we have had the second presenter for the industry representatives. Is Roger Briden here, from BioStar?

DR. BRIDEN: Thank you. I am Roger Briden. I am the Chairman of the Infectious Diseases Subcommittee for AMDM, and that is the Association of Medical Diagnostics Manufacturers, sometimes called the Association of Medical Diagnostic Manufacturers, as it was quoted in the initial agenda.

We represent some 50 companies of sizes varying from small startup companies to major manufacturers. All of the companies that we represent are involved in the manufacture of diagnostic products. We distribute those types of products and, combined, we conduct over one billion dollars in combined annual sales of these types of products.

We are both pleased and disappointed with today's panel meeting. Disappointed with FDA as it was our understanding that there would be opportunities to discuss with FDA the issues that were to be presented to the panel so that we might provide the most appropriate input. This did not happen in this case. We applaud the issuance of the examples and questions package. However, its lateness in coming out is a problem that needs to be addressed. With that said, we certainly look forward to putting an early communication process in place with FDA for future meetings.

We are pleased for the opportunity to participate with FDA in getting your input to this issue, and to working with FDA in understanding the ramifications and working out an approach that satisfies FDA's needs as well as the needs of the consumers and of the manufacturers.

I will restrict my comments to the micro area. We all realize that the old laboratory standards, as well as new assay technologies, have their own shortcomings as well

as benefits. The established standard for micro is culture. It is our understanding that current practice is to require culture data to establish sensitivity and specificity, and to not allow claims of superiority to the culture standard.

Manufacturers are limited to claims and data presentation of performance versus the currently accepted gold standard. Discrepant resolution is allowed, however, performance cannot be recalculated. The resolved results can be commented on in text. If the resolve method is run on all samples, then the manufacturer can display results versus that resolution method.

This policy presents a problem for new technology when such technology outperforms an existing but old gold standard. Let me provide an example of the issue. An assay has performance that exceeds routine culture. However, as routine culture is the gold standard, the manufacturer is not allowed to claim performance which it really has. This also results in labeling requiring things such as culture backup of all negatives, and in this type of a situation it would appear illogical to backup a test with one that is less sensitive than the original test.

Medical and patient treatment standards change as well as laboratory standards. Laboratory standards should be paramount in product reviews. However, practitioner use

standards should remain in the hands of the user. Labeling should not be so restrictive that it impinges on the user's options.

Relative to laboratory standards, we have seen the laboratory standards required by FDA change with time and to become combined standards using multiple methods, for example, enhanced gold standards. When FDA uses a new gold standard during product review and prior methods have been held to a different standard, this presents the problem of an uneven playing field for the manufacturers. Examples of changes are blood agar culture of strep A going to broth culture; TSA agar culture moving to lim broth for strep B; and cell culture going to cell culture combined with DFA for Chlamydia; and probe assays are now entering into that picture also. It is costly to manufacturers to generate new data and it is an administrative nightmare for FDA to get all manufacturers up to the new standard in unison.

As new technology emerges, we find old standards are no longer state-of-the-art. We are faced with a policy that doesn't allow claims of superiority to the old standard. We are, thus, faced with comparison tables that have significant numbers of discrepant results, the main area of concern often being potential false positives. We can run discrepant samples on another technology or method

as a resolution method. However, there will be some percentage of discrepant from that technology also. So, where do we stop?

Such resolution could continue through several iterations. Consider this approach: Evaluate the test product and the old standard, both versus the new standard. One can now assess the relative performance of the test versus the old standard while still establishing the performance of the test versus the new standard. One can claim more sensitive than the old standard but in a relative sense only. The labeling would contain the performance of both the test versus the new standard, as well as the old standard versus the new standard. This allows the evolution of the laboratory, or gold standard, in a controlled fashion. It has some down sides, in particular, the associated cost and administrative problems of implementation. It also requires extensive education of consumers as performance numbers are sometimes dramatically different, depending on how advanced that new standard is.

There are a variety of consumers and they use a variety of technologies. In practice, the healthcare delivery system in the U.S.A. is composed of many heterogeneous organizations, each potentially having different medical, clinical and laboratory standards.

In order to serve our customers best, manufacturers should have the flexibility to establish performance of a new product versus a variety of cleared technologies. If the manufacturer chooses the platinum standard and shows the performance exceeds the old standard, then they should be allowed to say that they are more sensitive than that old standard. The appropriate display, using a truth-in-labeling approach, of such performance would allow the individual user in the various settings to better match what they are using, or what they are most comfortable with, with the information that is presented in the package insert.

Another thought for consideration is whether there should be a gold standard reference lab to which the performance of labs involved in generating clinical study data are compared or certified. This perhaps could help in normalizing performance data versus a given gold standard. This approach is not without risks, and those risks may outweigh the potential benefits.

As we work through the issues of how to measure and report the performance of assays using new technologies, certain basics need to be kept in mind. First, manufacturers need to know what laboratory or clinical data needs to be gathered and how we can present it to our

customers.

Second, there needs to be a level playing field for claims and our ability to compete in a competitive marketplace.

Third, because in reality there exists a variety of customers, those being the healthcare system's providers using a variety of technologies across this group, the approach recommended to resolve performance questions relative to the standard method must incorporate a judicious amount of flexibility. That flexibility needs to accommodate today and into the future the evolving medical practice and laboratory standards.

Because of the significant complexity or the issues we are talking about today, we would encourage that a truth-in-labeling approach may be the most sane way to attack this problem. Thank you for your time.

DR. THRUPP: Do any panel members have questions of the two presenters?

DR. NIPPER: I have one. Mr. Briden, you are already in the back of the room, but I wanted to ask you, you used the words "gold standard" and then you used another word, "platinum standard." And I don't understand what you mean by that and maybe you could help me with that.

DR. BRIDEN: Okay. Sometimes I use the words

"tarnished brass standard" as a replacement for "gold standard." When I used "platinum standard" I just meant a higher level, perceived to be more technologically advanced, more sensitive. So it is the next step in a standard. If you have a gold standard, then what is the next better standard? That is what I meant by "platinum."

DR. NIPPER: Is your "platinum" standard one that has been cleared for marketing by the FDA, or is it a research standard? In other words, where are you mining your platinum?

[Laughter]

DR. BRIDEN: In this case it could be either. It sort of depends on the technology. My first reaction is that normally it would be first-level cleared standards, or cleared technology, but I would not limit it to that.

DR. NIPPER: Thank you.

DR. CHARACHE: Also for Mr. Briden. You obviously covered a great deal of material in a very condensed form. So, I would be curious as to whether we can receive a copy of the material that you put forward.

But I wasn't clear about one thing. I understand the need that you expressed early on for an evolving ability to assess new products as techniques change or opportunities change. I wasn't as clear about why if the culture

technology evolves it should not also be used as a standard of comparison, for example, the use of broth cultures for group A strep, if it turns out that there is good data now to suggest that that is a more reliable way of knowing that the organism is present or not present. So, I was having trouble resolving these two different concepts of a changing data base where new information becomes available that improves upon the way we were able to assess products in the past.

DR. BRIDEN: Yes, and I don't think I mean to imply that culture could not evolve. The issues with culture become largely technique. The ability of laboratories to do culture can vary significantly, depending on their ability. If those parameters which affect that were to evolve so that it made the reliability of culture results better, then I would say, yes, that would be an evolution also.

DR. THRUPP: Are there any other people in the audience that would like to ask a question of the presenters? Dr. Edelstein, did you have a question?

DR. EDELSTEIN: I don't know whether I am mirroring a theme expressed before but, for Dr. Ivor, the question of how to establish a better gold standard is one that I can't quite grasp. You said, for example, that some

of your assays are more sensitive than the current gold standard. But then the question is how is your newer comparative assay validated? In a sense, it is almost a circular argument unless you can demonstrate clinical data, and then it is also a very difficult issue of how can we define clinical disease in the absence of laboratory tests. It is very difficult in modern medicine to do that.

MS. IVOR: You have two questions, as I understand it. You are asking about the validation of the alternate amplification test --

DR. EDELSTEIN: Yes.

MS. IVOR: And the second one is how can you make a clinical diagnosis without lab data.

DR. EDELSTEIN: That is correct.

MS. IVOR: First, I think Roxanne referred to this or directed an answer toward this initially, there is no clinical clearance process or FDA clearance process for an alternate amplification test. It is as though the manufacturer is developing two tests at the same time. The one that is undergoing clinical evaluation can be validated by its sister product, which is a test directed at a different genetic sequence of the same target RNA. So, that is a research means of verifying that the target is there; that your product is there.

DR. EDELSTEIN: Do you think that that second validation test can be accepted unconditionally as a gold standard though?

MS. IVOR: I think it is additional information that helps to validated those samples that appear to be false positive. Bringing into account your second question, certainly the clinical symptomatology of the patient is additional information that should be considered when looking at whether or not this is, in fact, a true positive specimen; whether that patient has disease.

DR. EDELSTEIN: But if you are only applying it to a certain patient population, then you have a number of biases that are introduced. You may, in fact, have some circular definitions. Can you tell me that the reference methodology, the reference amplification methodology is completely independent of the performance of your new test methodology?

MS. IVOR: Completely independent as far as?

DR. EDELSTEIN: Is it possible that the results move in the same way because of similarities in technology?

MS. IVOR: I am not sure I am clear on your question. The two tests will not detect the same sites, the same RNA sequences.

DR. EDELSTEIN: But is it possible that through

some other means there could be dependence or lack of independence? For example, the presence of something that is not an analyte yet affects both assays in a similar fashion?

MS. IVOR: You are talking about interfering factors, and so on. That can happen with any test, culture, immunoassay --

DR. EDELSTEIN: That is correct.

MS. IVOR: Yes.

DR. EDELSTEIN: Because if that were the case, then you would bias the discrepant analysis, if they move in the same direction. One question I might have for you is have complete data sets been reanalyzed using these reference methodologies? Not just the discrepant samples but the complete data sets?

MS. IVOR: Non-discrepant?

DR. EDELSTEIN: Yes.

MS. IVOR: Not to my knowledge, no.

DR. EDELSTEIN: I think that information might be very useful to see if what has been modeled in a recent journal article that you mentioned--whether, in fact, that modeling is accurate.

MS. IVOR: As far as the validation of the alternate assay, the manufacturer is responsible for

validating that in-house, and as far as running that on negative specimens, positive specimens, and so on, that is part of the workup.

DR. EDELSTEIN: Thank you.

DR. O'LEARY: This is aimed at both of the presenters, and it is a little thought experiment. The question is, you are trying to claim substantial equivalence between, say, a nucleic acid amplification assay and a culture assay, except that you are saying that the nucleic acid amplification assay is better. Now, I take a case that I believe had, say, influenza from 1918 and I put it in culture and I culture it out very nicely but I put it in a DNA amplification assay and I detect it very well. I have done a great job of detecting, in this case, mRNA sequences but I have not done a very good job of detecting viable organism.

And one of the questions here is are we trying to ask in what situation we can label oranges as apples? Are we saying that we are detecting disease or an organism? Should we just be labeling for detection of nucleic acid sequences of Chlamydia or the organism de jour? Is this really a substantial equivalence question at all? Aren't we really talking about quite different things? So, is there an industry response?

DR. THRUPP: Mr. Briden is rising to respond to your challenge.

DR. BRIDEN: One of the sections I was going to bring up but I didn't was that the difference we see is that culture detects, as you said, viability of the organisms. Most of the newer technologies we see are detecting either the presence, either current or past, of a marker for that organism. That is basically the way the technology is evolving.

It is a good question, and I think it is one that tends to be organism dependent as to whether or not its immediate viability is an issue. That gets, in my opinion, more into clinical practice and what the significance of this is because you are really looking at a surrogate marker for the organism. But I think from an industry position we look at the truth-in-labeling approach to say "this is what it detects." Exactly how you use that information then is the question that the practitioner needs to feel comfortable with.

DR. THRUPP: I think that is part of the information that we are being asked to comment on today, what is the meaning of the apparently non-viable organism in the target populations for which the package insert would direct the test to be used. Dr. Ivor?

MS. IVOR: The actual claim that is made is for detection of that ribosomal RNA; it is not for the disease state.

DR. THRUPP: I think Dr. Gates had a comment.

DR. GATES: I was just underlining what they said. I think it has always been a kind of fundamental question in terms of in vitro diagnostics, is it a surrogate for the disease state or is it the detection of an analyte? I think in most cases it has to be the detection of an analyte because there are a lot of other things--specimen handling, the lab technique for, one, running the test and, two, making a diagnosis. It also speaks for the fact that if that is the case, you have to test it against a predicate because in that case--you can't test it against a disease state per se. You have to test it in terms of how it works with other predicates, both involved in testing for some particular analyte.

DR. THRUPP: I would throw out one generic comment. We have heard comments as to the FDA's need, the manufacturers requirements and needs and the laboratory users, and I would just make the apple pie comment that we are really all directing our efforts to two areas, the healthcare in the public as well as healthcare of individuals. That should be the bottom line of where all of

these discussions are going. Dr. Charache?

DR. CHARACHE: I think my question is directly along that line. I am not exactly clear, if you could help me understand exactly what your perspective would be in terms of providing information on how to use a given test, and what the FDA should require in order to permit that kind of information to be provided both to those who perform the test in the laboratory and to their constituents, who are the clinicians using the test.

Specifically with the example of, let's say, Chlamydia, if one reports the presence of RNA what information should be provided and what documentation should there be that a given result which measures such a marker is or is not correlated with a disease or a given predictive value of the disease? Certainly, we are very keenly aware of the fact that the diagnostic capacities have long since exceeded the ability of the users to know how to use them. And, part of our major job is to educate the user on how to use the information. I could go further with my favorite organism de jour, which is adenovirus, whose presence may be shed as a colonizer, as a latent virus, or may be associated with disease. What kind of information would you think is appropriate for the FDA to ask of you so that the laboratorian can advise the clinician on how to use the data

if it is not a viable organism?

DR. BRIDEN: Excellent question. I wish I had a really good answer for it. From an industry's perspective, or as a manufacturer, we are certainly interested in saying what information we do need in our package insert so that the user of the product can use it appropriately. One of the difficulties we face is that the practice of medicine changes. It evolves with time as the standards we have been talking about. So, at any one given point in time there will be practitioners out there who are using diagnostic assays as an aid in what they are doing, and that is how they are basically touted as, as an aid in the diagnosis. They are using them for different purposes, and this then gets into the realm of practice of medicine and something that we really would have great difficulty in doing other than trying to provide them a tool to use in their studies or investigations.

The information which should be requested, in my opinion, should be, again, the issue of truth-in-labeling. What should be required is a clear statement of exactly what was the data that is presented; how was it generated and where did it come from. Then the practitioner, I would hope, would be able to use that information to best understand is this assay, and the way I am using it,

appropriate. I think it becomes very difficult to answer the question and, hopefully, we will look to the panel or to other presenters perhaps to say, as a practitioner in the field, what is it that you get out of the package insert. I think someone answered that question as, "well, once in a while I read it." That presents a problem in itself. So, if they pay little attention to it, it becomes very difficult to decide what really is the most valuable information in there. That is where I would come back to the data that was presented, how was it generated and how were the calculations and things done.

DR. CHARACHE: What kind of guidance would you feel was the responsibility of the manufacturer to provide? We have already said the average family practitioner in a community hospital can't make those decisions on his own. They are depending on someone else to guide them.

DR. BRIDEN: Again, I would go back to basic truth-in-labeling: Here is the data; here is the population in which it was generated; here is how it was generated. As manufacturers, we are not in the business of making medical practice decisions. We would need to stay out of that arena.

DR. THRUPP: That issue can be brought back to the floor later. There are still a couple of other comments. I

think if you could hold them, we will have an opportunity for the panel, obviously, to discuss further. We have four more presentations. At the end of those, hopefully, we will have a couple of minutes to ask for any other comment from the audience that is with us today.

We have responses from two professional groups or associations, Dr. Schachter's response and Dr. Miller's. They were unable to be here in person so Freddie Poole will read the letters that they have provided us.

Professional Group Responses

MS. POOLE: In the interest of time, I am not going to read every line in the letters but I will read the important points that were raised by Drs. Schachter and Miller.

Dr. Schachter states that for many years Chlamydia culture was considered the gold standard but even in the best hands reproducibility and multiple specimen testing suggested that culture was only 75-85% sensitive. With the introduction of antigen detection methods, a proportion of antigen-positive specimens were culture negative. When nucleic acid probes were introduced, they were more sensitive than antigen detection but still less sensitive than culture.

The problem as he sees it is, number one,

amplified nucleic acid technology introduced, such as PCR and then LCR, than TME, with those more positives were detected than with culture. Antigen detection methods confirm only a fraction of nucleic acid positive samples that were culture negative.

Number two, discrepant analysis using alternate probes, primers or other DNA amplification tests confirms a great majority of nucleic acid positive, culture-negative specimens, as true positives.

Three, the goal of discrepant analysis is to determine whether the initially positive assay is a true positive or a false positive.

Number four, discrepant analysis, attacked by statisticians as a biased approach to evaluation of diagnostic tests, overstates the sensitivity and specificity of the amplified nucleic acid test.

Number five, a crucial issue evolving from the statistical criticism is the precision that confirmatory tests are invalid unless all specimens are tested by the confirmatory test.

Number six, there is a subset of specimens that are positive only by nucleic acid amplification and cannot be confirmed by culture or antigen detection methods, but are confirmable by other nucleic acid amplification tests or

use of different targets.

Number seven, he states that discrepant analysis is certainly not perfect. Now that three nucleic acid amplification tests are on the market, there have been a number of evaluations where different specimens from the same patients were processed by all three assays. There is very little difference in the results obtained by discrepant analysis versus the use of multiple assays on all specimens. Furthermore, the use of antigen detection methods on all specimens in parallel have been shown not to generate new positives.

On balance, results obtained with discrepant analysis seem to be quite comparable to those obtained by any other method, and discrepant analysis is far less expensive and labor intensive. He states that it still seems to be a useful method for evaluating diagnostic tests and should not be abandoned until actual data are generated for better methodologic approaches. To abandon this approach because of criticism, without having a validated alternative, is not productive.

Dr. Miller states that under ideal conditions using a perfect test for resolution of discrepant samples, discrepant analysis leads to an overestimation of sensitivity and specificity. Under less than ideal

conditions, when an imperfect test is used to resolve discrepant samples the bias may be considerably larger than under ideal conditions but may also be smaller.

Alternative procedures can and should be used rather than discrepant analysis. I believe discrepant analysis in its current form is not an acceptable procedure for the evaluation of new diagnostic tests. The bias, although small in some circumstances, can be huge. Given that the goal of discrepant analysis under ideal conditions is a biased result, I cannot recommend its use. Thank you.

DR. THRUPP: Since Dr. Schachter and Dr. Miller aren't here, we can't direct questions to them. Let's move right on. Our next presenter will be Dr. Timothy Green, from CDC. He serves the AIDS, STD and TB research lab divisions. Dr. Green?

CDC Presentations

DR. GREEN: Good afternoon. My name is Timothy Green. I am with the Division of AIDS, STD and TB Laboratory Research of the National Center for Infectious Diseases.

[Slide]

I will briefly describe an evaluation of bias in diagnostic test sensitivity and specificity estimates computed by discrepant analysis that appears in the

February, 1998 issue of the Journal of Clinical Microbiology. In doing so, I wish to make it clear that the views expressed here are those of the authors and do not represent an official position of the Centers for Disease Control and Prevention.

[Slide]

The context of this work is the determination of sensitivities and specificities for nucleic acid amplification tests used to diagnose Chlamydia trachomatis infection. Typically, these determinations have been made using a two-test design with cell culture as the reference standard. The use of cell culture as a reference standard is based on the presumed high, perhaps even perfect, specificity resulting from the reliance on visual identification of specifically stained inclusion bodies.

It has long been recognized, however, that culture sensitivity is considerably lower and that it varies substantially among laboratories, making culture an imperfect reference standard at best.

On the other hand, it is biologically plausible that nucleic acid amplification tests have much higher sensitivity than culture while retaining very high specificity. Indeed, evaluations of these tests, using culture as a reference standard, typically yield a

substantial number of culture-negative, amplification test-positive specimens.

Believing that most such specimens come from infected persons, many investigators have adopted a practice of applying one or more additional tests to these specimens to determine whether the positive amplification test result can be confirmed.

[Slide]

This table illustrates such an experiment where LCR is used as an example of a nucleic acid amplification test and LCR MOMP, an alternate target test, is used as an example of a confirmation test. Each specimen is tested by both culture and LCR, and culture-negative, LCR-positive specimens are subjected to a MOMP test. The culture-based estimates use the culture test to classify a person as infected or infected, while the discrepant analysis-based estimates classify a person as infected when either the culture test is positive or both the LCR and MOMP tests are positive.

Since discrepant analysis removes the culture-negative, LCR-positive, MOMP-test positive specimens from the denominator of the culture-based LCR specificity estimates and adds them to both the numerator and the denominator of the culture-based LCR sensitivity estimate,

the discrepant analysis-based estimates of both LCR sensitivity and LCR specificity are always greater than or equal to the culture-based estimates. This does not necessarily mean, however, that discrepant analysis-based estimates are more biased than culture-based estimates.

To assess the accuracy of published estimates of amplification test sensitivity and specificity, we compared the bias in estimates based on discrepant analysis with that in estimates based on culture. Comparisons were made over realistic ranges of values for culture sensitivity and specificity, LCR sensitivity and specificity, and the prevalence of *C. trachomatis* infection in the study population as indicated on the next table.

[Slide]

We used the generally accepted culture specificity value of 100% but also allowed for a slight degradation of this value, with the amount of degradation increasing with increasing prevalence of infection. In addition, we included the case in which LCR sensitivity is the same for culture-positive as for culture-negative specimens, as well as the case in which LCR sensitivity is moderately higher for culture-positive than for culture-negative specimens. We set the MOMP test sensitivity and specificity for culture-negative, LCR-positive specimens to the mid-point of

the ranges used for overall LCR sensitivity and specificity, but also examined the effect of varying the MOMP test sensitivity and specificity values over the broad ranges used for the initial LCR test. In selecting values for the remaining test performance characteristics and prevalence of infection, we attempted to reflect both what was included in the manufacturers' package inserts and what has been published in peer-reviewed articles by independent investigators.

[Slide]

This graph and the one following show the bias in percentage points in both culture-based and discrepant analysis-based estimates. The red is culture based, the blue is discrepant analysis based. The X axis is the actual LCR specificity. The Y axis is the bias in the specificity estimate in percentage points.

The culture-based estimate of LCR specificity is biased downward throughout the indicated range. This bias increases as prevalence of infection increases and as culture sensitivity decreases, that is, as more specimens from infected persons are culture negative.

On the other hand, the bias in the discrepant analysis-based estimate of LCR specificity may be upward or downward, but what bias exists is small and is generally

less than that of the culture-based estimate. This is because removing LCR positive specimens from the denominator of the LCR specificity estimate, even using an imperfect confirmation test, largely eliminates the underestimation of LCR specificity caused by culture-negative specimens from infected persons.

Furthermore, other biases, particularly the overestimation caused by not removing similarly misclassified LCR-negative specimens from both the numerator and the denominator of the estimate are negligible.

[Slide]

The effect of discrepant analysis on estimates of LCR sensitivity is more complicated. The ideal estimate of LCR sensitivity would be based exclusively on specimens from infected persons and would include all such specimens. Such an estimate would be unbiased. If culture specificity is 100%, the culture-based estimate of LCR sensitivity is based exclusively on specimens from infected persons but only includes specimens that are culture positive. If LCR is equally sensitive for culture-positive and culture-negative specimens, including only the culture-positive specimens does not introduce any bias. Therefore, the culture-based estimate of LCR sensitivity remains unbiased.

If, instead, LCR is more sensitive for

culture-positive than for culture-negative specimens, including only the culture-positive specimens causes the culture-based estimate to be biased upward. Since adding culture-negative, LCR-positive specimens to both the numerator and the denominator of the culture-based estimate increases the estimate, discrepant analysis either creates or increases upward bias.

In short, if LCR sensitivity for culture-positive specimens is either equal to or greater than that for culture-negative specimens, there exists a number, albeit a small number, of culture-negative, LCR-negative specimens from infected persons. Failing to add these specimens to the denominator of the culture-based estimate of LCR sensitivity is detrimental to the accuracy of the discrepant analysis-based estimate.

Conversely, the direction of any bias in LCR sensitivity estimates is less predictable if culture is even slightly less than 100% specific. The presence of as few as 1-4 culture-positive test results per 1,000 specimens from infected persons introduces a substantial downward bias in the culture-based estimate of LCR sensitivity. This bias is downward because most of these culture-positive specimens will be LCR negative and, thus, included only in the denominator of the culture-based estimate. It is

substantial to the extent that applying even a very small false-positive culture rate to the large number of specimens from uninfected persons in low prevalence settings produces a substantial number of culture-positive specimens compared to the much smaller number of specimens from infected persons. In this case, discrepant analysis may improve the culture-based estimate of LCR sensitivity by introducing an upward bias that offsets the downward bias caused by culture-positive specimens from uninfected persons. The estimate may, thus, be reasonably accurate but only to the extent that competing biases not fully taken into account by discrepant analysis cancel each other out. This seems a poor justification for using discrepant analysis to estimate LCR sensitivity.

[Slide]

In conclusion, the bias in estimates of nucleic acid amplification test specificity based on discrepant analysis is acceptably small, and is generally less than that in estimates based on culture. However, the accuracy of discrepant analysis-based estimates of amplification test sensitivity depends critically on whether culture specificity equals or is slightly less than 100%, and it is affected by competing biases that are not fully taken into account by discrepant analysis. Thank you.

DR. THRUPP: Thank you, Dr. Green. Let's move on to Dr. Alula Hadgu, who is not from the Division of Sexually Transmitted Devices --

[Laughter]

-- but I think he probably comes from the Division of Sexually Transmitted Diseases.

[Slide]

DR. HAGDU: The purpose of this presentation is to provide you a short overview of the published literature on discrepant analysis. It is also to share with you my opinion about discrepant analysis. This is only my opinion and is not necessarily the opinion of CDC.

[Slide]

As you know, discrepant analysis is an attempt to identify the truly positive patients that cell culture testing misses. In discrepant analysis the apparent false-positive individuals--individuals are positive by the new test and negative by the imperfect old standard and are subject to additional ancillary testing, generally by the same amplification test or by a similar amplification test.

[Slide]

So, in terms of references, the first critical review of discrepant analysis I published in Lancet, in 1996, and article called "The Discrepancy in Discrepant

Analysis."

A more mathematical and algebraic version of my Lancet paper was published in Statistics in Medicine, in June, 1997, and there is an interesting set of letters, actually, to the editor also published in the Lancet, in 1996.

[Slide]

My conclusions are that sensitivity and specificity estimates obtained by discrepant analysis are biased, misleading and dangerous. Even if we use a perfect test to resolve the discrepant results, even if we resolve cell b by code, discrepant analysis estimates are still biased in favor of the new test. Thus, this technique should not be adopted in evaluating the performance of a diagnostic test. Those are my conclusions.

[Slide]

In July, 1997 both the editors of the Lancet and Statistics in Medicine commissioned a commentary on my work on discrepant analysis. This commentary was entrusted to a fellow called Jorgen Hilden who is a professor of biostatistics at the Department of Biostatistics in Denmark. He is also a medical doctor. Dr. Hilden was chosen because he was independent of all of us and he was an expert in diagnostic testing issues.

Dr. Hilden said discrepant analysis is a ploy to exaggerate claims of performance. He characterized discrepant analysis as based on faulty logic and fallacious statistical arguments. In the unpublished version of his paper he called discrepant analysis a damned lie --

[Laughter]

[Slide]

Perhaps the most prominent researcher in diagnostic testing issues, ladies and gentlemen, is a fellow called Colin Begg, from the Sloan-Kettering Institute. Colin Begg said the following: Discrepant analysis is fundamentally unscientific. It is conceptually and logically flawed. I suspect that no article that focused purely on statistical biases could persuade me this is a valid scientific approach. I agree.

[Slide]

More essentially, Drs. Green, Black and Johnson published an article in the Journal of Clinical Microbiology. Their conclusion is the following: Nucleic acid amplification test specificity based on discrepant analysis is small and generally less than that in estimates based on culture.

Incidentally, Dr. Miller has a forthcoming article in the Journal of Clinical Epidemiology, next month. The

sgg

title of the article is "Bias in Discrepant Analysis When Two Wrongs Don't Make it Right." His conclusion is the following, even when a perfect test is used to resolve discrepant results, Miller demonstrated the presence of a substantial bias associated with the use of discrepant analysis estimates. That is completely contradicting the conclusions of Green et al.

[Slide]

I also have a forthcoming article called "Patching Up Discrepant Analysis." The conclusion of that is the following: Using actual discrepant analysis studies, resolution by the MOMP test, and using the assumptions of Green et al.--I want to stress that the bias associated with discrepant analysis estimates is large and it is more than that in estimates based on culture. Again, the opposite of Green's paper.

[Slide]

Here are two examples where I actually took published papers, one on the cervix and one from the package insert. This curve shows you the bias in the culture based- and the discrepant analysis-based estimates of LCR specificity. Here values were obtained using the assumptions of Green et al. You can see that the DNA-based specificity is much further from the truth than the

culture-based. The bias is huge. Again, this is using their own assumptions.

[Slide]

All these things are good. You know, you can use tables, you can use figures, you can use mathematics, you can use algebra to look at things, but the most important thing in science is to go to first principles. Physics has first principles. Mathematics has first principles. Probability theory has first principles. Even governments have first principles. The first principle in diagnostic testing is that the new test should not be used in the determination of the true disease status. In discrepant analysis, ladies and gentlemen, the definition of true disease status is based in part on the outcome of the new test under investigation, the plasma-based DNA, and its own sister test, the MOMP-based DNA. This is analogous to the new test being the judge and the defendant simultaneously in a court of law, and that is not consistent with mathematics, physics or even common sense. That is all I have to say.

DR. THRUPP: There will be more discussion yet to come, but I think we can take two minutes to ask if there are any comments from the audience before we go to the committee discussion. Does anyone else from the audience that is not a presenter or not a Committee member wish to

sgg

make a comment? Gee, this whole crowd is silent? I can't believe it! Yes, could you identify yourself and come to the microphone? Thank you.

DR. WENG: I have a few questions to ask. When --

DR. THRUPP: Excuse me, could you give your name and affiliation?

DR. WENG: I am Teng Weng. I work for my boss, Greg Campbell. My remark is inspired by him. The problems of discrepancy analysis or calculation of sensitivity and specificity are mostly from confusion in issues of definition. We have to make a logical distinction between sensitivity and predictive values. Sensitivity is defined from disease to the test result. Disease is logically prior. Predictive value is from the test result to disease. So the test result is logically posterior. When you have a true gold standard, error free, no error at all, only in that case, given the disease rate, do you have a definition of sensitivity. If you don't have a gold standard, a true gold standard, everything you calculate, even what the test result is for the other test, and in that sense you are calculating some form of predictive value and not sensitivity at all, and depending on how you use it, it may be wrong too.

DR. THRUPP: We have to move to the committee

discussion time.

DR. WENG: Okay. I do have a mathematical formula to show why the discrepancy analysis went wrong.

DR. THRUPP: Thank you. I think it would be fair to let a couple of our panel members who had questions or comments in the previous discussion--I think Dr. Hammerschlag and Dr. Taube had their hands up before.

DR. HAMMERSCHLAG: Well, I hope they are not out of context, but in some cases, looking at this not only as a researcher but also a clinician, it is almost akin to trying to decide how many angels are dancing on the end of a pin. Are we ever going to get 100% sensitivity and specificity? I don't think so. Is it necessary that we really achieve that? I don't think so. The issue is to know the performance of the test and have some minimum acceptable standards for this performance, and to educate people as to the limitations of the tests and when they should be used appropriately.

Unfortunately, clinicians are not the ones who are making any decisions today on what tests are being used. The decisions are being made by laboratories, usually by laboratorians--which is a term I have heard and I hate that; it is like I tell my medical students and residents not to "med speak"--and often the decisions are really not even

sgg

being made on performance but often purely on economics, and sometimes the decisions on what tests are being used are actually being made by people who are not even physicians, not even Ph.D.'s but administrators. Certainly, we have dealt with certain HMOs and they are dictating what laboratories the test may be going to and quite often the physician has really no idea.

The technology I think has run away with most people in practice not understanding the technology. The laboratory is a black box. They put in the slip; an answer comes back and they have no idea what is going on in the box. And to say that they will read the package insert, they don't read the package insert. They have never seen the package insert. Actually, many people in the laboratories themselves do not read the package inserts. I have a ton of anecdotes I could give you.

The issue about disease and how tests are being used, when we do tests for Chlamydia, they are not really being used to diagnose the disease state because most people with genital chlamydial infection are asymptomatic. You cannot tell by looking. Maybe 75% of women with chlamydial infection in the endocervix are not going to have any signs or symptoms on physical examination that can be directly attributable to say this is Chlamydia. We screen them

sgg

because we know that if we identify the organism--and in culture, if you identify Chlamydia in somebody they are infected by definition. You are dealing with an obligate intracellular parasite and there are certain sequelae that we are trying to prevent in terms of, certainly in women, ascending infection, pelvic inflammatory disease and all the morbidity that could be associated with that, prevention of infection in infants, etc., etc. So there is no clinical state; it can't be a clinical diagnosis. The issue is does the presence of DNA have the same validity as the presence of viable organisms for the risk of developing these sequelae later on.

Tuberculosis, however, is a horse of an entirely different color because we generally screen people for exposure with PPD and then they are prophylaxed. In that case, they do come in with a clinical syndrome. I find it very difficult, however, to probably come up with certain clinical criteria or scoring system for these patients, especially when one is dealing with immunosuppressed patients.

On the other hand, for something like pertussis it might be very, very possible to come up with a clinical scoring system on presentation of patients with a syndrome and evaluating a test in patients who are asymptomatic, who

sgg

don't have the syndrome, that you might come up with something. So, I think it is going to be very important to understand these limitations of how the tests are basically being used, and to accept that we are never going to reach nirvana in this case. The point is how to educate people to the appropriate use of the test, and maybe there does have to be some truth-in-labeling to say, you know, these are your probabilities. And, I think I have gone on long enough.

DR. THRUPP: I would add one corollary to Dr. Hammerschlag's comments. It is true that the package insert is never seen even necessarily by the laboratory director, and the FDA may have limited means to require that these things be disclosed. Nevertheless, in the computer reporting era there is a much better opportunity--and how to require it is another issue--for a select two or three lines of limitations or caveats to be appended to the value or the result that is put into patients' charts. Perhaps in the long range we should pay more attention to making a recommendation as to what limitations or interpretive comments should be appended to the report. Dr. Taube?

DR. TAUBE: Actually, my question related to a number of the other questions and comments and, basically, it comes back to the idea of the fact that there is

information in the package insert about what you are detecting. The question was raised, I think by Dr. Charache, about what is the relationship between the marker and the disease. The question I think that everybody is struggling with is what kind of tests can you do; what kind of evidence can you develop that establishes the relationship between the marker and the disease? And it is that kind of information that the clinician needs to know in making the clinical decision about what the test positive or negative triggers in terms of the care of the patient.

DR. GUTMAN: I just want to comment on this because FDA is not entirely naive and we don't make the assumption that everyone reads our package insert from cover to cover, and that it is distributed to all medical students and practitioners. However, we take the package inserts very seriously. We do assume that they are an important resource. The one thing I certainly agree with that the industry said is that truth-in-labeling is something that we strive to work with manufacturers to attain whether they are read or not. We believe that is the right thing to do, and we think they probably should be read. Maybe we are not doing a good enough job about making sure how important they are.

One of the deliberations we are asking for, and

Dr. Taube is right, is what tools are available and in what ways to express what we want to say, and we are going to make the assumption that physicians are intelligent enough to be able to read what is put in the labeling and put that in some kind of context. That may in some cases be a dangerous assumption. That is, nonetheless, the assumption that we make.

DR. O'LEARY: I have sort of a question again, and I hate to throw more questions out to muddy the water but in some cases, like the nucleic acid technologies and considering culture as a predicate maybe the mistake is in considering it as a predicate. You know, in one former life I was an analytical chemist, and the first analytical chemist that came up with an assay for nickel had to do it on the basis of theory and detailed knowledge of what the reaction was, and how that was going to go, but not on a previously available assay for nickel.

Similarly, offsetting errors can be nice but they are problematic. Once upon a time quantum calculations on small molecules got done by something called semi-empirical methods because there were some offsetting errors that made up for an error in the ab initio methods. But as the ab initio methods got better we got rid of the semi-empirical methods which were mathematically flawed. So, maybe we have

to not forget first principles in theory and rely too much on the predicate devices. I think if the theory is sound--it is absolutely required that the theory be sound, and if the theory isn't sound, and I would have to distance myself from the opinion that Dr. Schachter had in his letter--no amount of apparent practical significance will make up for a flawed theory.

DR. THRUPP: Did you have your hand up, Dr. Nipper?

DR. NIPPER: Yes. I want to comment on Dr. Hammerschlag's statement about the diagnosis of Chlamydia not necessarily being a clinical one and being a laboratory-based diagnosis. It underscores what I try to teach the medical students at Creighton when we go through the lectures on this particular topic, and even a history and a physical have a predictive value, a sensitivity and a specificity.

Therefore, I think that when we talk about platinum standards, gold standards, alloy standards, etc. I think that the kinds of standards that have to be applied to a particular comparison need to be case based. We have two analytical chemists sitting elbow to elbow; he went on to something better but I stayed one. Dr. O'Leary and I happen to agree with a lot of these things that are being said by

each other. There was a national reference system in clinical laboratories that was being developed so that we could have better analytical standards. I hear a lot of the problems with these techniques being described today that can be traced back to imperfect analytical techniques being used as comparative methods.

I also hear an undertone and, Dr. Gutman, I hope I am not leading us into the swamp here. If some of these methods don't compare well to a predicate device, should they be in 510(k)s at all? In other words, are we really in the right regulatory mode when we start talking about trying to find some analytical technique in the corner that we can drag out and use as a predicate device? Am I asking the wrong question here?

DR. GUTMAN: I think so. I think you are straying into a regulatory issue. That is very interesting because if it doesn't match the predicate there are two possibilities. One is it doesn't match the predicate and it shouldn't be on the market, and the other is if it doesn't match the predicate it is still a damned good device and it should come to the market through a PMA process --

DR. NIPPER: Yes.

DR. GUTMAN: -- but I would much prefer that you concentrate on the scientific issues and let us obsess over

sgg

the color of the jacket. If you would like we could reconvene a panel later and we can talk about that --

[Laughter]

-- but all you have to do is solve the answers to the questions that Sharon posed and we will be perfectly happy.

DR. THRUPP: A quickie?

DR. HANSEN: Yes, a real quickie and in complement to what Steve was saying. As I have alluded to earlier, the reason that we need your advice and recommendations very much is because we are not going to be looking at nearly the kinds of submissions we used to. We are going to be concentrating on higher risk devices. Our concerns are how they best can be characterized. So, we are not worried about the old tests, essentially low risk. We are worried and concerned about those tests that we will be looking at, either as a PMA or 510(k)s, which we believe will be a major sole determinant to aid in diagnosis, monitoring, etc.

DR. NIPPER: That is why I asked my question.

DR. HANSEN: Right, Henry, and I was with you all the way.

DR. THRUPP: I would like to throw out one comment in response to Dr. Hammerschlag's and Dr. Nipper's comment about the Chlamydia specifically. It would be, I think, a

sgg

mistake to feel, in a broad sense, that the Chlamydia test is laboratory based in its total sense. What you meant is that there is a laboratory test which is required because the issue can't be solved on clinical examination. However, the value of that result has been established by clinical studies which have shown that the positive culture, in the absence of symptoms and/or nucleic acid or the presence of antigen in some form or another, does correlate with later clinical problems. So, in this instance there is a clinical basis for establishing the relevance of the laboratory test, which is not necessarily true with all the other areas where we might be looking at pieces of DNA that might have been residual from something that happened 30 years ago and may or may not be relevant today or in the future.

DR. HAMMERSCHLAG: Also, when you find the evidence of the Chlamydia, I mean it is going to initiate treatment and intervention which is relatively low risk. I think we are very well aware of the clinical sequelae of what is happening, but then it becomes an issue, again, as to what kind of levels of accuracy are we really going to strive for if we understand the limitations of the test in the population that we are using it in and the realization that it is never going to be 100% sensitive and specific under any circumstances. You know, how far can we go on

with that? What is acceptable? In some populations, for instance like sexually active adolescent girls who win the award for having more Chlamydia than anybody else probably, you know, we realize that reinfection is a common issue and you are going to have to do repeated testing so that, hopefully, you will capture most of that population.

On the other hand, sometimes the test gets used forensically and we can accept that it can't be used under those circumstances, especially in evaluating it in prepubertal children. So, this is the information that has to get out there, as to when it is appropriate to use the tests and what they mean, and when it is inappropriate to use the tests. And, again, how far do we have to go to prove that it is perfect or close to perfect?

DR. THRUPP: I think we have to move to looking at the questions in a more specific fashion. Dr. Hansen, do you want to lead us into the questions? Whether or not we will get answers is problematic.

DR. HANSEN: Let's start out with example 1.

DR. THRUPP: If everyone has examples a, b, and c in hand, what claims should be allowed for each of the examples a, b and c? This was the ELISA assay in example a. It was tested on a relatively small sample with a high prevalence of the condition. What claims can be allowed?

DR. HANSEN: Keeping in mind what the intended use of the device is. Another way of asking is can those claims be validated? In other words, the device is intended to aid in the diagnosis of something.

DR. THRUPP: In this case, individual patients with a disease. And we assume that the criteria for the disease have been well defined and the FDA has been over those definitions --

DR. HANSEN: No, I would not make that assumption.

DR. THRUPP: The definitions of the disease, it seems to me, are a basic assumption before you can evaluate anything.

DR. PEPE: I think that the disease is defined by the predicate. Isn't that correct?

DR. THRUPP: In this example we are assuming --

DR. PEPE: In this particular example I think that is the case.

DR. THRUPP: Okay. Yes, Dr. Todd?

DR. PEPE: There are a lot of holes in this data set.

DR. TODD: It is very difficult. I mean, a primary principle should be that we are trying to use these tests to make a diagnosis and, therefore, realistically we need to begin to think about some kind of outcomes data so

sgg

that as our technology continues to improve and we can change and still know that we are dealing with real data. You can't just keep comparing it to previous tests that may be flawed.

DR. PEPE: Right, but on the basis of the information that is here, this is how the new test compares with the predicate.

DR. TODD: Right, but I guess I am saying that what I would want included in the package insert is information about the predicate --

DR. HANSEN: And how the samples were characterized?

DR. TODD: How the sensitivity and specificity of the predicate was determined.

DR. HANSEN: That information was not available, nor was the case definition for the 23 specimens from patients supposedly with the disease. There was nothing in the submission to say how that disease was diagnosed.

DR. EDELSTEIN: I am awfully confused about the presentation of data in this table. Where it says predicate device, should I read diseases, disease population? Or, do these actually compare two different devices?

DR. HANSEN: They actually compare two different devices with the specified populations as expressed in the

sgg

explanation, Paul.

DR. EDELSTEIN: Okay.

DR. HANSEN: The positive represents 23 specimens from "patients with the disease," and 25 represent blood specimens or serum specimens from blood donors, and the predicate tested those specimens and the new device tested specimens, and those were the results.

DR. EDELSTEIN: You are saying that the predicate device had 100% clinical sensitivity and 100% clinical specificity.

DR. HANSEN: I am not saying that it has 100% clinical specificity. This is relative sensitivity.

DR. GUTMAN: Again, you can change the parameters any way you want, and we would rather not be leading so we would rather you changed the parameters. But Dr. Pepe had this right, this is essentially based on definition of disease by the predicate, not by the patient population. You may be horrified or delighted by that. That is irrelevant to me. What is important to me is based on that, what --

DR. HANSEN: What can you claim.

DR. GUTMAN: What can we put in the package insert based on this kind of performance data? So I wish, in retrospect, that we had made this a quantitative assay and I

sgg

wish the analyte had been sodium or hemoglobin, but in a qualitative way I can't figure out how to do it with sodium or hemoglobin. We are talking about an old analyte that has been around for a long time. You know, with hemoglobin we are not asking that it be demonstrated that low values are affiliated with anemia any more; we ask for an analytical base. So, try and think of this as a qualitative hemoglobin.

DR. CHARACHE: Personally, the only way I could begin to answer that question would be to translate this into a hypothesis. So, I am going to hypothesize now. I know nothing about this presentation and I am going to say, okay, this is Lyme disease. I have 23 patients who have been diagnosed as having Lyme disease and 25 blood bank donors. What can I say from these results?

What I would say is I want to know more about what kind of Lyme disease they had. Are these people with new disease and a skin lesion? Are these people who have arthropathy? Are these people who have CNS disease? Because I could see very great differences between early and late diagnoses and security of diagnoses. How did they decide, if it is arthritis, that they had Lyme disease? Did they decide on the basis of the predicate test that it was Lyme disease when we know that the early predicate test had

sgg

a high cross-reactivity with helicobacter pylori, with syphilis and a whole bunch of other things.

So I would say that I need more information. If this is supposed to be a model, and I just chose something to model with, I need to know more about the clinical diagnosis. I need to know if this is going to be used, what the population is that I should be using this test in, and I will give you one bias which I will only say once, but I have a great problem with this percent overall agreement of 93% because you could make that a very high agreement if you had 900 blood bank donors in there and it wouldn't have anything to do with why you are doing the test, which is to know if a given infection is present.

DR. THRUPP: I agree with you, Pat. We would like to know all the criteria on which the predicate device was established as being a predicate device. It would probably be reasonable to recommend that there certainly be in the package insert data on what established the predicate device, and at least some summary for the new one too. But, as I understand it, we are being asked for the assumption that there is a predicate device that has been assumed, based on whatever data could be reviewed, how does the new device compare with the predicate device without getting into the clinical discussion.

DR. GUTMAN: You are at the edges of what we might like to ask but I am not sure we could legally ask. Maybe we could convert this into a drug and say this is a cocaine test. Would that help?

DR. THRUPP: Okay, but we have to make the assumption that the predicate device is real under whatever circumstances, whatever background data is available. We are not going to get through all these questions if we don't move on. Dr. O'Leary has been waiting.

DR. O'LEARY: I am going to answer four questions of my own first. On example a, I wouldn't say much of anything, and the reason I wouldn't say much of anything is because the number of cases is small and my guess is you would have to have a 40% difference to reject the null hypothesis that the two devices are equivalent. That is a best guess on the usual sample size for a contingency table. So you have to define early on, before you do the study, how much of a difference you are going to accept and get your sample sizes worked up according to your best estimate of the prevalence in the population and, you know, do a proper sample size estimation to begin with.

It is hard to make anything out of any of the examples if you don't know anything about the study design, and the study design was not put forth to begin with in a

sgg

reasonable way.

Similarly, with the sensitivity and specificity data you have problems of all sorts but, at the very least, I would think in the end, however one decides to get that stuff out, you ought to provide some confidence intervals because the confidence intervals here are going to be astronomical--my best guess, not an engineering number of a physics type number.

The last comment is just with regard to the business of using a discrepant resolution and how you would deal with that part. I wouldn't. I think the discrepant resolution doesn't have a sound mathematical basis and my personal view is that the agency should bar its use in the future.

DR. THRUPP: Could I throw that back to you? Let's suppose that this was not just cocaine, which is around a lot, but suppose this was a rare disease that the manufacturer went to great lengths to collect these 23 known cases from all over the world and had a very valuable, difficult to collect set of positive samples in a very rare disease. How would you respond to that?

DR. O'LEARY: Well, it creates problems and it is something to think about. What I didn't mention is that false positives and false negatives have a cost to them of

sgg

one sort or another, and the decision you make depends a little bit on what you think that cost is going to be. If treating somebody that doesn't have a condition or disease is cost free, then that false positive doesn't make a whole lot of difference. But if I have a test for cancer that, on the one hand, is going to kill somebody in three months if untreated but in which the treatment has a 25% mortality associated with it, then I am probably going to want to have a rather tighter set of confidence intervals for making my diagnostic test based on that cost that is associated with screwing up in either direction. And, I think it is really impossible to generalize the numbers in a simple statistical test. I really think you have to think in a broader decision analysis mode in order to make sense of that kind of decision on a predicate device because, again, for a high risk device I am going to want those confidence intervals to be real, real tight, and I am going to want the theory behind it to be real, real sound by comparison with the case where maybe it is not going to make a big difference.

DR. THRUPP: You raise the issue of confidence intervals, and throughout these data presentations and data sets--maybe Dr. Gutman could comment on one issue that has not been brought up, namely, is there data or should data be required for the predicate on replicability in given

laboratory devices because that is going to affect your confidence intervals, aside from the sampling errors and the biologic errors?

DR. GUTMAN: Yes, these have been simplified but generally for both quantitative and qualitative tests we would look for repeatability. If it was a quantitative test we would actually look for various components of variation.

DR. THRUPP: But this issue is part of statistical variability that is not really addressed --

DR. GUTMAN: Well, we are trying to simplify --

DR. THRUPP: Right.

DR. GUTMAN: -- and there is an interesting internal discussion about what term to use when you are comparing a new to an old device. This is perhaps a particularly brilliant or particularly poor example, but one of the issues is whether we should be using the term relative or comparative or estimated sensitivity and specificity, or whether we should move away from that and use the term overall agreement. I have heard at least one brick thrown at that. Or, whether we should use some other term, or whether we should use no term or at all, or whether there is some way to express whatever is going on here.

DR. HORTIN: I think for the first question that was addressed to us, we are kind of getting lost in the

sgg

specifics a little bit. Whether there are 10 samples or 10,000 samples, or whether we use discriminant analysis, or what predicate device, I mean, basically the claim is going to boil down to that the device is going to be used either in the diagnosis of a disease or for the detection of infection.

I think the point that we may be able to address here is whether any claim beyond that would be allowed, whether it would be providing better sensitivity than, say, culture methods or any other claim beyond that. That would probably not be beneficial because I think you should simply refer people to the data later in the package insert.

I think in practice the first issue here, the claim is going to be very simple and generic, and whether they did a study doing 10 samples, whatever predicate device they used, or whether they did 10,000 samples the claim is basically probably going to be the same and we should really probably not allow expansion on that too much, and should basically refer people to the data as to more specifics in terms of the performance characteristics.

I guess the only question about a nucleic acid claim would be should it be worded specifically as nucleic acid detection or whether it should be detection of the organism. I think we are probably getting a little bit lost

sgg

in the specifics and the details of the examples.

DR. THRUPP: Well, that is true. Nevertheless, Dr. O'Leary has made a very cogent point that is real. The numbers in this example, and we are trying to give a response on this example, are small enough that if you did apply a statistical predictability you are going to have a broad range there. Before we could have the recommendation say that the claim should be that this test is 87% sensitive and 100% specific in comparison with the predicate, should we require that there be a statistical range placed on those statements in a package insert or an allowable approval? Yes, Dr. Kroll?

DR. KROLL: I am going to reiterate what Dr. O'Leary commented on. I think it is very important here to look at the numbers. You can't use a percentage if your numbers don't add up to at least over 100, and they clearly don't here. That is really erroneous. That is really misleading. Even if you put it in the package insert, it is misleading.

The second thing is that we can only really refer this to what we find for the predicate. If it is a test for hemoglobin, it is a test for hemoglobin not a test for a disease.

DR. THRUPP: Well, that is what we have already

sgg

discussed. We would want it to be stated quite clearly.

DR. KROLL: But, I mean, it has to be listed as a test, not just in the package insert but in what they call it, what they name it, because sometimes these things are given out as a test for a certain disease or condition --

DR. THRUPP: Those are important points to be transmitted, hopefully, to the package insert and, hopefully, to how it is reported.

I think we had better try to move on if we have at least a little bit of a consensus.

DR. NIPPER: How about an answer to question 1 in one or two sentences? Question 1a, I don't think you can claim anything --

[Laughter]

-- I think you need to go back and ask for more testing from an appropriate population. On 1b, I think you can claim substantial equivalence, although I think it is debatable depending on what the device is. On 3, I think you can claim that it is a device for the diagnosis of, and you put in whatever disease it is because you are doing a clinical evaluation. It is not the best one I would like to see but it is there.

I have a problem with using percent overall agreement. That is the efficiency of the test and it should

be stated as such because that is a defined term. But I have a problem with all three of these examples because I don't think they are well designed.

DR. THRUPP: The examples may not be --

DR. NIPPER: The examples are fine, but I mean the studies are not well designed.

DR. THRUPP: The examples are problematic but the FDA has to face these kind of examples so we are trying to give them guidance on what to do.

DR. NIPPER: Right.

DR. THRUPP: Your conclusion for example a) was that we don't let them say anything. But suppose this is a rare disease and they are not going to get more than 23 samples of positives, would it be reasonable to allow a conclusion that it is equivalent to the predicate device with a statistical range based on the numbers that are available?

DR. NIPPER: In your example I don't think it is substantially equivalent to the predicate device because 3/23 don't agree. In your rare disease you should be able to do better than that analytically. I am speaking as a chemist now. In other words, I think the test needs to be improved. I don't think it ought to be allowed.

DR. THRUPP: That makes the assumption that it is

the type of test--well, okay.

DR. KARMEN: I read these questions as being separate and independent. The first one, I would like to give the people who are submitting this the benefit of the doubt and say that if they only studied this small number of people it is because they had something that was a highly specific assay, chemically specific, and that they could determine something from this small number.

When I tried to imagine what this could be, I thought of HCG as a pregnancy test. And, if somebody was going to give me a test that has less analytical sensitivity than the predicate test, I would want it to be a hell of a lot easier. So, if this were a strep test that could be done in 30 seconds perhaps, very easily by anybody, and they could show that it had the same specificity as compared to normal--presumably they do this with women but not necessarily because we have found positive pregnancy tests that were approved by the FDA that were positive in men --

[Laughter]

-- because of hama present. I would have said that this has less sensitivity than the predicate test and I would define the clinical situation in which you could use a less sensitive test, and what cautions you would have to have about repeating it in another week to be sure. This

sgg

will become positive after so many days.

Then the other questions I think are similar. But from the next two examples here I lost my HCG because it seems if this were the same product being described by the next two, it is neither as sensitive nor as specific as the predicate, and you would have to say that and I would want to know why these folks are presenting this as a useful test. I will quit there.

DR. THRUPP: Perhaps at this point that is enough comment about this. So we will move on to -- did you have one more on this?

DR. EDELSTEIN: I apologize, but I find it very difficult to deal with these examples in abstract because there are certain principles that we have discussed that we have to keep in mind. One is, what is the actual disease we are discussing and what are the implications for diagnosis and treatment? And, what we are willing to accept as acceptable performance depends greatly on that, as we have discussed.

The other issue has to deal with what is an acceptable control population. Without knowing the disease prevalence and implications of diagnosis or misdiagnosis it is impossible to know what an appropriate control population might be.

DR. THRUPP: Well, you just answered very nicely question 2. You know, the characteristics are going to be dependent on the populations and the disease or the predisposition to the disease, whatever it is that it is targeting. So, it is hard to come up with a generic statement that is going to cover all situations. Steve, I don't see how we can come up with a specific generic recommendation that is going to be independent of the specific populations or specific disease.

DR. GUTMAN: Yes, I am not sure. Sharon, do you have a suggestion?

DR. HANSEN: We tried to make these very generic rather than specific because we are well aware, as you have stated to us, that each disease has its own entity. ELISA technology is a very popular technology. If you chose what, let's say, is high risk disease, what we are trying to get at is the approach, and I think in many ways you have answered that. If you don't have the information you don't know how to make recommendations. Certainly, you have said that.

DR. EDELSTEIN: I don't agree--I mean, it has to be a specific analyte because the same principles are here. If this analyte were a tumor marker for breast carcinoma or bladder carcinoma--think of it that way as an example for b

and c.

DR. O'LEARY: But I think I disagree on that, and the reason is because if I compare that and the decision there versus an adenovirus test intended for somebody who comes into the physician's office with an upper respiratory infection, then I am probably talking about relatively different outcomes, depending on how things work, and so I am probably going to work with different criteria.

That sort of brings you into this question of number three, which is how and when should specimens from health donors and normals be used. The answer is probably only in a screening test, and one of the real issues you come up with there is when you have doped this and you have used that as your normal base and then seeded patients from something else. That reminds me of some of the early study sets on some of the PAP smear screening stuff. You know, if you take that and then you compound it with the discrepant resolution business you are actually amplify the potential biases in the discrepant resolution. A quick run through my head on the sensitivity of that at least suggests that that is probably true. So I think you really have to look at the specific application and, again, the consequences of an incorrect decision to make sense of it.

DR. THRUPP: I think we should move on. Number 4

sgg

has been mentioned a couple of times, which is perhaps semantics but they are important. What about the terms we use, relative sensitivity or specificity? Do we like comparative sensitivity and specificity or estimated? I think we have heard some comment that the overall agreement can be a misleading figure depending on the prevalence of the condition. But what about the terms relative, comparative or estimated? Anybody have any thoughts?

DR. PEPE: I guess I prefer the term sensitivity for something in particular, and to define what that something in particular is.

DR. THRUPP: Well, we would assume that you are talking about the sensitivity of X compared to whatever the predicate.

DR. PEPE: These are all estimated sensitivities.

DR. THRUPP: They are all estimates, exactly. But what term? We have heard from the FDA a couple of questions as to what terms should be used.

DR PEPE: Well, I don't like the term relative sensitivity and the reason is because it suggests that you are comparing the sensitivity of test A for a gold standard versus test B for a gold standard.

DR. THRUPP: And that is not what we are doing.

DR. PEPE: No, I think what we are doing is we are

sgg

getting the sensitivity of the new test for the predicate test result.

DR. THRUPP: So, do you like an adjective in there at all? Estimated? Dr. Nipper?

DR. NIPPER: I want to tell you that I have always been disappointed that sensitivity and specificity were used to describe these statistics because analytical sensitivity and analytical specificity were around long before these became used, and they get confused in my lab and in my mind all the time. So, if we are going to call it a different term let's think of something other than sensitivity and specificity, define it and get on with it. Rather than trying to modify these misnomers in the second place, let's move on. Let's let the statisticians help us with that and let's move on with it. I wouldn't dignify it by modifying it with relative, estimated, or any of that stuff.

DR. THRUPP: Yes, I don't think we are going to be able to come up with completely new terms at this point but I am not sure that the adjective adds anything to the terms as they are being used at this point in time.

DR. NIPPER: No.

DR. TODD: But I think it is an important point that if you are going to use sensitivity or specificity you want to say whether it is diagnostic sensitivity and

specificity or analytic.

DR. THRUPP: Right.

DR. NIPPER: If you say it is analytical and use it for these concepts, you are going to get it confused with describing the analytical technique.

DR. TODD: Right.

DR. NIPPER: So, that is why you either ought to stick with clinical or diagnostic, or something, or make sure you define your terms. Otherwise, you are going to get over into the technique area.

DR. THRUPP: There is a second part to this question, number four. Should the test reports to the clinician reflect the nature of the performance characteristics and, if so, how? I am not sure what "nature" means.

DR. NIPPER: No.

DR. THRUPP: Steve, do you want to comment on what we are after there?

DR. HANSEN: Steve is looking at me. I would agree with the distinction, Henry, that you just tried to make because that is confusing. That is of great concern to us, the definitions, and that everybody understands things the same way.

What we were really trying to ask is in example a,

which you have essentially said you can't say anything about, example b, there is a targeted population. You don't like the term "relative." Example c, however, was based on a clinical study. So, there are subtle differences.

DR. THRUPP: Well, it is more than subtle. Those are big differences and they should be defined in the recommendations and the approvals.

DR. NIPPER: But should you put that in the computer that goes back to the clinician with the test results? Is that what that question means?

DR. HANSEN: Yes.

DR. THRUPP: Yes, if you can. I must admit that is sometimes easier said than done. Much as I was advocating this, when our lab computer people came out with long paragraphs it became a morass that nobody read either.

DR. GUTMAN: Well, you can recommend that as not practical.

DR. O'LEARY: Isn't that really the responsibility of the laboratory to see what makes sense, because you should be validating some performance characteristics in your own laboratory and that is the thing that is most important to run back to your clinicians. I would say we toss this back to the labs.

DR. THRUPP: But that is a different issue in

sgg

terms of quality control and replicability within the lab. I think we are talking about the use of a test that is out there on the market.

DR. O'LEARY: We clinically validate in our lab as well. I am sorry, but we can't afford to do things just because --

DR. THRUPP: You have to rely on FDA's decision, huh?

DR. O'LEARY: Well, no, but the FDA looks at--you know, there is one set of information that is very useful predicate information but the situation in which something is approved for use does not necessarily precisely replicate the clinical situation of our patient base. So, we have to understand it in terms of our patient base.

DR. THRUPP: That is true, but in terms of your point, I would submit that 99% of laboratories out there are not going to be able to go through reestablishing predicate validity, etc., in their own populations and they are going to have to take the package insert or the data in the literature in order to validly apply their test, given that they do the appropriate quality controls, etc.

DR. CHARACHE: I am going to say the test result reports to the clinician should not include the performance characteristics for two reasons. First, because the

clinicians won't understand what you are trying to tell them, particularly if you have something like that on every test that has been approved by the FDA. Secondly, they will get mad because you are using too much paper and you are wasting their time.

DR. TAUBE: I think that the information that needs to go back to the clinician is the information that the clinician needs to make a decision. So, it should tell the clinician what was exactly evaluated so if it was nucleic acid versus culture, you know, live, viable organisms, that should be on the information. Then there should be some indication of what that means. So, finding nucleic acid indicates that there is some remnant of the organism but not necessarily that it is viable. But as in the case that Dr. Hammerschlag brought up before, there should be some indication that you should treat or not treat based on the disease.

DR. THRUPP: You are going a little too far afield I think --

DR. TAUBE: All right, I retract that but some indication of what the professional organization suggests.

DR. THRUPP: Even that is going to be difficult, but your point is well taken. Dr. Gates?

DR. GATES: I am just a little confused because a

sgg

little bit ago we were talking about needing a lot of information in terms of the disease, in terms of the patient population and in terms of interpreting the data we are seeing, and now it sounds like we are saying that we shouldn't let the person that is actually making the diagnosis for treating their patient--we shouldn't tell that person what that data is. So, I mean, do we need the data?

DR. EDELSTEIN: We can't interpret it without knowing the clinical information which only the clinician has.

DR. NIPPER: I think the solution to your problem, Dr. Gates, is that I have found it very useful in cases I have worked to have this kind of information available to the using physician in the laboratory handbook rather than on the chart. There are certain few comments that we put on the chart with each test result but those are generally held to a minimum, and data like this about test result reports and how the performance is characterized belong in the laboratory handbook, which I hope most physicians would read or even know where one is.

DR. HAMMERSCHLAG: Yes, I think this is the point. I mean, we have to educate and this may be one way of educating. I was also thinking that maybe we need the "medical letter." You know, they always talk about a new

sgg

drug. Trovafloxacin is just out; grepafloxacin--same thing, a new diagnostic test and an assessment of its performance. I don't think it should be on the slip but there is an education role that somebody has to play because technology is getting much more complex.

Dr. O'Leary talked about clinically validating. One problem I have noticed, and the laboratory people have actually complained to me that when they get requests for tests, even though there might be a space on the request form for the clinical presentation, often none of this data is ever provided, not to mention often sending inappropriate samples. Again, you know, they may have a perfect test but if it is used inappropriately it doesn't matter and then it goes right back to the real-world situation of clinicians where, again, the technology is running ahead of them.

DR. THRUPP: There are certain things that can be practical to do but we can't solve all the problems of the reporting.

We had better move on. I think on number five we have had kind of a consensus of a few comments here. As presented, are the resolution for discrepant results statistically and scientifically appropriate? We have heard a lot discussion and presentations about this issue. Does anybody want to add any more comments about discrepant

sgg

analysis?

DR. PEPE: Well, I would like to applaud Dr. Hadgu's comment that the gold standard should not be defined by the new test, and that is part and parcel of the discrepant resolution method that is applied, as it is applied right now. I wonder though if in spirit what the discrepant resolution is trying to do is trying to use, say, the culture and resolution test together to define--it may be bad, but a gold standard; to define disease as present if either one is positive. Maybe that might be a gold standard. Anyway, a gold standard needs to be defined in every case and it can't be defined by the test. One could imagine developing statistical methods to calculate a sensitivity and specificity that would be relevant to, say, either the culture or the resolution test being positive but discrepant resolution does not get at that either. It is just impossible to interpret.

DR. THRUPP: It would if there were a 100% sample of the discrepant reference test in addition to the predicate and the new test, but that is not practical in the real world necessarily unless there would be some circumstances where the FDA might decide, without guidance, that it were necessary.

DR. PEPE: I would like to hear back from

industry. It would be possible to estimate such a thing by just taking sub-samples from each of the four cells in the table.

DR. THRUPP: If you are dealing with low prevalence phenomena samples aren't going to solve the problem, I don't believe, statistically. I think Dr. O'Leary was next.

DR. O'LEARY: I actually was wondering if we could get Dr. Hagdu to make a comment on alternative approaches to the gold standard, perhaps based not on the new diagnostic test but multiple diagnostic tests. I think he has given some thought to this problem.

DR. HAGDU: Estimation of statistical performance indices in the presence of an imperfect gold standard is quite a difficult problem. No matter how you slice it, it is a very difficult problem.

There are two possible solutions to this. One is what you call back-calculation. You look at the relationship between the new test and the imperfect gold standard and you can back-calculate the sensitivity and the specificity. There is a cost to that. For example, if you want to estimate the specificity of the new test you can mathematically express precisely the specificity of the new test as a function of the cells a, b, c, d and what you

think is the sensitivity of the imperfect gold standard. So, one way of doing it is by back-calculation and mathematical adjustment once you have observed cells a, b, c and d and, of course, then you have to have some kind of estimate for the sensitivity of the imperfect gold standard.

The second is mathematical modeling, and statisticians have a lot of tricks. Sometimes they don't work. This is not a new problem. Statisticians have been working on this. There is a modeling technique called latent class models in which one can use several imperfect tests, three, four, five imperfect tests and concoct and create a gold standard out of three or four imperfect tests. It is difficult to explain this without mathematics, but conceptually what that is doing is the following: Imagine that this room is dark; there is no light. We can't identify each other. There is one light. It doesn't help us identify each other but if you continue to add imperfect lights there is going to be a point at which the imperfect lights provide us sufficient information to identify each other.

So, latent class models have actually been used to estimate sensitivity and specificity of new tests in the presence of no perfect gold standard. The problem with latent class models is that they come with an assumption, an

sgg

assumption of conditional independence, which is not the right case in many cases. So, like all statistical things that come with an assumption, that assumption isn't generally attainable.

Recently there has been work on latent class models with random effects in which that assumption of conditional independence is relaxed. There is a paper by Qu which was published in Biometrics where he actually does estimate sensitivity and specificity in the presence of imperfect gold standards using random effects.

I also have a paper coming in the Journal of the Royal Statistical Society. This has been accepted and should be published in five or six months. I used latent class models with random effects to estimate sensitivity and specificity of imperfect gold standards. Qu and I also have another paper coming out in the Journal of the American Statistical Association. That is coming in June, 1998, and that also uses the latent class model.

So, to answer your question, one is back-calculation after having observed that imperfect table of the new test in culture; and the other one is using mathematical modeling.

DR. PEPE: Well, a question I have about all of those approaches is how is the gold standard defined?

DR. HAGDU: Mathematics.

[Laughter]

DR. PEPE: That worries me.

DR. HAGDU: Of course it has to worry you; it should. As I said, there is no easy solution to this. There are rational techniques and rational approaches. Discrepant analysis is irrational. This is one of the rational techniques in which, as I said, you can use imperfect gold standards to concoct and create information. You are basically borrowing information from weak tests and creating information that actually mimics the truth, and there has been a lot of work on this.

DR. THRUPP: Does the panel have any other suggestions on the use of the discrepant analysis procedures, especially as have been applied in some of the recent examples? Dr. Charache?

DR. CHARACHE: Just one other thought. I have had problems with discrepant analysis as it has been done in the past, but particularly if one only looks at those boxes that represent discrepancies without concurrently looking at the remainder of the populations. The FA was used as an example of the resolution. There was one test we were involved with in which we discovered that the FA was measuring something different than either the ELISA or the particle technology

sgg

which we were comparing with each other. It was just coming up with a very different answer and we learned that by looking at the population as a whole.

So, I think it is doubly flawed when only the discrepant things are examined and then they go to augment the apparent sensitivity and specificity. We have seen this also when the clinical situation was used as the discrepant resolver as opposed to a laboratory test. I remember one example of this in which an investigator, in a published study, was looking at blood culture systems and for the discrepancies where the new method picked up 33 more isolates than the old one, they decided that it was true if the patient had a fever and looked septic but, of course, that is when you get a blood culture and it turned out that all 33 were the most common skin contaminants. But that was used as a marketing strategy for 15 years. So, I think that there are a lot of arms to this and I think had they looked at the same criteria for all 4 boxes it would have dissuaded them from using that as a discriminator.

DR. THRUPP: I think it is time for us to stay on schedule and take a 10-minute break. We will reconvene and we will try to be unequivocal about "equivocal" in example 2.

[Brief recess]

DR. THRUPP: Our audience is fast diminishing but, for those that are in the audience, at 4:30, hopefully, we will have a chance for further response from those who are not members of the panel if there are further comments.

Let's go on with addressing the questions, and there is a question in relation to example two. Dr. Gates?

DR. GATES: Just a general one in relation to number two and the idea of discrepancies.

DR. THRUPP: Going back to example two, where it says the very low prevalence disease where the new test is being proposed with an equivocal zone response, whereas the predicate device is black and white, so when you are dealing with this kind of a situation what is the appropriate presentation of the new assay's performance in this situation? Who would like to tackle the issue of equivocal? Nobody! Dr. Edelstein will tackle it.

DR. EDELSTEIN: Well, I think that a simplistic solution is to exclude the equivocal results from the analysis. That would be my suggestion.

DR. THRUPP: Any other?

DR. PEPE: I don't feel comfortable with that. I think that in this case it would make the test look pretty good, whereas there are 19 samples where you kind of don't know anything after having performed the test, and I think

that that is an important component that should be described.

DR. THRUPP: I would think that would especially be true if you are dealing with an example like this with a very low prevalence analyte. Throwing them out completely, I am not sure is an answer.

There are two phases or two levels of considering this. One, of course, is an application or the study to evaluate the new device and submission to the FDA, and then a second level is what should be in the package insert or how should the laboratory handle them? So, for these purposes I think, Steve, we should probably address the first issue of how is the FDA going to evaluate the studies on evaluating the new device. Dr. Pepe?

DR. PEPE: It states here that when there is a sample that is in the equivocal zone then the user should rerun the sample. I was disappointed that they didn't actually rerun these samples to see what information the test ultimately gives you using that protocol of rerunning the equivocal results. If you still get entirely equivocal results, then that means that it is not a very informative test for a substantial number of samples. Whereas, if you get unequivocal positive or negative results on the reruns, then at least you have a conclusion.

DR. NIPPER: I was thinking during the break about this issue, and I think I would like to build on what Dr. Pepe was saying, and that is that from a clinical standpoint it is unsatisfactory to have an equivocal test and leave it at that. I think that if it is important to test in the first place, especially with the neonatal marker, then you need to come to a clinical conclusion and that clinical conclusion cannot be equivocal. So, therefore, the test should be evaluated on the system that is proposed in order to reach a conclusion, and it should be labeled as such, that it is only usable in that system of screening and then confirmation even by getting a second specimen or having it referred for confirmatory testing, and so forth. So, I don't think you should calculate sensitivity and specificity on equivocal results. I think you calculate it on the final result that you get when you do the testing to its appropriate conclusion.

DR. KROLL: Yes, I agree with Dr. Nipper. I think what they should do is compare how the previous predicate test and how the new test compare against the confirmatory results, and they shouldn't worry at all about sensitivity or specificity, or anything like that, and from that they can decide whether or not the new test does a better job.

DR. THRUPP: So if, as suggested, the test is

sgg

repeated, whether it be a replicate of the same sample or even with a new sample, and you still get equivocal results or a result in what has been determined to be a borderline quantitative range, you then ignore it in calculating specificity and sensitivity, etc? Just throw them out? What do you do with them?

DR. PEPE: I would be inclined to include them and define as positive those who are truly positive so that your sensitivity is not as large as it could be if you had conclusive information, and for your specificity, similarly, those equivocal results would not be regarded as negative and so your specificity is also hurt by unequivocal results, as I think it should be.

DR. THRUPP: So, you calculate it both ways. Dr. Nipper?

DR. NIPPER: I am unsure about your question, but I think if you define in the package insert, which is prefaced by appropriate studies, that two equivocals end up being in the negative and that is validated clinically, then you no longer have equivocals. In other words, you have to get rid of the equivocals before you calculate anything and if you define two equivocals as negative and that is supported clinically, it is no longer equivocal. If you define two equivocals as positive, or if you define one

equivocal or two equivocal as the gold standard method at a state laboratory and then you use that state laboratory results as the final, then you have your sensitivity and specificity based on that. I just think you have to get to the point where you have a 2X2 table --

DR. PEPE: That is right.

DR. NIPPER: -- Not 2X3 table.

DR. PEPE: Right, but you can't exclude them entirely from the analysis.

DR. NIPPER: No, you have to take them to the final conclusion. Whatever clinical decision is made or whatever testing decision is made as reportable, then that is what you use to calculate the sensitivity and specificity.

DR. THRUPP: You are backing into the same issue that we have dealt with in extenso today, namely, you are retesting a small subset of your samples and essentially instead of doing discrepant you are doing equivocal reanalysis and you are getting --

DR. NIPPER: I totally disagree with you. I do this routinely in toxicology testing where we take a presumptive positives, take them to the GC mass spec. We don't report anything as positive unless it is confirmed by the GC mass spec. We don't even talk about the sensitivity

and specificity of the screening test as a whole, we talk about the whole system. So, the system is not new. It is what works well and, you are right, we are retesting only a subset but that works clinically for the needs of the testing population. So, this is not reinventing anything.

DR. HANSEN: Could I ask a question, please?

Henry, let's assume it is a screening test --

DR. NIPPER: Yes.

DR. HANSEN: -- what if it is supposed to be a diagnostic test and there is an equivocal zone?

DR. NIPPER: Then in order to get useful information you have to have some system of reporting that in an acceptable way to the clinical community. If there is no acceptable way to report that to the clinical community so that appropriate clinical action can be taken, you have to handle that within the lab to order a retest. We do that with neonatal testing in my clinical lab, back in Omaha. If we didn't get an appropriate sample or if we got an equivocal test with a couple of the screens we did on neonates, we then brought the mothers back in for a second specimen and then we reported to the state health department. So, I think you have to have a system by which you operate so that you produce a clinically useful result and then you can calculate sensitivity and specificity of

sgg

your testing system. We did that with biotenadase. We are one of the few states in the nation to test for that particular rare disease, and we had a confirmatory test for the biotenadase deficiency.

DR. THRUPP: I think we have a number of questions relating to example three that we should move on to. Are there any other comments on the quantitative equivocal? If not, let's go to questions from example three.

A lot of discussion has already taken place about these issues, but let's see if we can come back to a little bit more nitty-gritty answers. Number one, for future evaluations of any Chlamydia device and other diagnostics where more sensitive technologies are being used, or at least allegedly more sensitive technologies are being used, should the gold standard be redefined to include testing by a nucleic acid amplification test, that is commercially available or well described and its performance validated, or should culture continue to be used as the gold standard? Dr. Hortin?

DR. HORTIN: I think there are really two points in the evaluation. One is that actually in terms of the information that is important and useful to generate out of this, it is important for people to get information about a new test compared to the most commonly used procedure that

sgg

is used diagnostically now. Whether you call that a gold standard or not, I think that may be a misnomer, but I think in terms of one aspect of the data that should be provided for a test, it is hopefully to look around in terms of common application for what is used in terms of diagnostic testing now and to generate some information about that because in terms of figuring out in terms of information for comparative purposes, how this test is going to compare to the old one that people have been using maybe for 5 or 10 years, they need that. If they move immediately to kind of a new, improved gold standard, you are kind of making a leap and you leave everybody behind. So, that is kind of one point.

The second point I think is probably a little bit more difficult issue. You know, we have brought up some points about whether you are measuring viable organisms or whether you are perhaps measuring non-viable organisms. I think it probably is useful to generate data about that as well in terms of reference data, but I think we have been trying to think in terms of perhaps absolute truth here but I think the starting point is to know how we are comparing with what the existing state of practice is in terms of the practitioner and also the laboratory. I think that we don't want to forget about that.

DR. THRUPP: Could I throw out just two scenarios, getting back perhaps to two ideal worlds as a place to start for an answer? Let's suppose the data were available in scenario number one, that an NAA test has been thoroughly looked at in relationship to cultures, including sequential studies and a significant number of NAA-positive, culture-negative tests were found in people who had had Chlamydia disease five years ago and they have been followed and they have had no further problems since then and, clearly, this was a little bit of nucleic acid left that didn't have clinical meaning. In that case, the clinical data would say that the false-positive test isn't truly a false positive and you want to retain the culture as the gold standard.

Alternatively, for scenario number two, if there was extensive clinical data showing that yes, indeed, that person from five years ago that had Chlamydia then and has a little bit of nucleic acid left in the sample and it looks like a false positive but that person is, indeed, subject to recurrence or to infertility and it has real clinical meaning, and there are studies treating that patient to show that they had less problems subsequently, then you would say that the NAA test should be the new gold standard. Let me throw those two scenarios to Dr. Hammerschlag. Is the real

sgg

world somewhere in between?

DR. HAMMERSCHLAG: No, not really. I think there are some studies that show that using Amplicor PCR or the LCR that the DNA does disappear after treatment. There are some recent data from Hopkins demonstrating that it might persist for as long as nine days but it does eventually go away. So, you would expect that if a person was diagnosed by one of these assays and treated, five years later it should be due to their probably having reacquired it. So, I don't think that would be necessarily the issue.

To me, the possibility, now that we have alternate tests available that use different technologies so that we are not caught in the bind, for instance of using the MOMP-based primers for both the LCR or the PCR where you may be trapped by the problem that both of those tests, although they are using different targets, use the same technology. You have alternative technology now with TMA and the possibility of using one of these tests in parallel is very attractive in evaluating a new test. But I still think I would like to see culture in there as well. Frankly, if they were all run in triplicate or in parallel, then I think we wouldn't even have to deal with the issue of discrepant analysis.

DR. THRUPP: I didn't quite hear you come to the

conclusion as to whether you felt that the culture should still be the gold standard, or are you suggesting that the gold standard should be revised?

DR. HAMMERSCHLAG: I think we have enough data to suggest that culture probably cannot be used as the gold standard. Certainly, it can be used as a gold standard probably for specificity but for sensitivity it is clearly not a gold standard. But that can vary.

I mean, the sensitivity of finding Chlamydia has been estimated in the endocervix, if it is present, to range anywhere from 60-80%. I think if you are looking at conjunctival specimens in babies with Chlamydia conjunctivitis you might be approaching 100% due to the factor that it is a more easily accessible site and there is much more organism there and fewer things that are going to interfere with the culture. Then you have just the technical problems of culture.

There was a recent paper by Ned Hook and Mitchell Pate that was in Sexually Transmitted Diseases about a year or two ago that looked at the variation in culture technique from lab to lab. You know, we are not dealing with a standardized method. It is not like isolating E. coli, putting it on the API strips, etc. There is a bit of an art to it, and there are all sorts of little glitches in doing

sgg

Chlamydia culture. But I think there are a certain number of labs that have a known performance, and that may need to be considered very much as to where these tests were done. I hate to say this but in some of the previous papers we have read, it is clear that some of the manufacturers, in a rather cynical and calculated way, have definitely selected laboratories that probably did not have optimal culture methods.

I would still like to see culture in there with the understanding that it is not a gold standard but I think we do need to modify it.

DR. THRUPP: Perhaps we should go right on to question 2 in example 3, in the interest of time. We are expecting submissions using NAA technology and other molecular-based methods, obviously, for detection of a variety of other etiologic agents, and Chlamydia, Neisseria and Mycobacterium are just examples of those that are coming along. What do we recommend to establish the performance characteristics of such devices when, example a), no culture methods are available or are impractical even when there are culture methods?

We had a comment earlier about MTB in the sense that there has been discussion and concern about the use of "clinical" diagnosis of TB when you don't have a culture.

So, that has been addressed a little bit. There are problems and I think a lot of us are a little reluctant to rely solely on these clinical characteristics. But there is culture for TB but what about if there is not a good culture method available? Dr. Charache?

DR. CHARACHE: I am just going back to our criteria for home brews and thinking of viruses that there are no culture techniques for--parvovirus B19 where it is impractical, rotavirus which is impractical. There are many settings in which I think a well-defined clinical parameter is a reasonable thing to use, and there may be nothing else that makes as much sense. I think it is a case of defining the clinical parameters which you are going to use that will define your disease entity. The two that I mentioned happened to work out pretty well. But I think that then you test your new test against it and see whether it picks up only patients with those syndromes which are consistent with those diseases. So, I think there is a population of entities in which you use the clinical.

This, I think, can also be helpful when you have a discrepancy. Again, I am thinking of things like varicella virus, clinical versus the culture technique for varicella versus the PCR for varicella, and PCR is much more sensitive than the culture technique but the tie-breaker is a

well-defined clinical study.

DR. HAMMERSCHLAG: Chicken pox, fortunately, is something for which you can usually make a good clinical diagnosis. I think where it may be important is when you have unusual situations at unusual sites, such as cerebrospinal fluid where getting a rapid diagnosis would be extremely attractive. The same thing I think would apply to the whole issue of enteroviruses and CCSF, and also getting back to TB. I think the data I have seen on the use of nucleic acid amplification tests for MTB and sputa seems to correlate that it is really no better than being smear positive. Am I right or wrong on that?

DR. EDELSTEIN: I think it is better.

DR. HAMMERSCHLAG: A little bit better?

DR. HANSEN: Yes, it is better.

DR. HAMMERSCHLAG: I guess I haven't reviewed that literature, but I think that one very important use would be in CSF. The data I have seen on that has not been that great.

DR. THRUPP: The question up front is, given that circumstance where, let's say, culture methods aren't going to work, how are you going to establish performance characteristics for such devices? It is based on clinical syndromes that you can define.

There are two possible levels for this question again. One could address what needs to be done on the developmental data for submission for approval on a research type basis as opposed to what should be recommended in the clinical package insert for clinical applications. It may be that there are circumstances where a difficult and impractical, very expensive--but a research or reference gold standard would be a special culture method with co-culturing or something that might be doable for the original validation studies but wouldn't be practical for clinical applications. So, you could still have two levels of a so-called gold standard. How should positive and negative predictive values be estimated for such devices?

DR. NIPPER: You have to choose your test population well and then do the study, otherwise why would we want to do the test if the positive predictive value is not any good? We need to know that, and that is generally not just about this particular issue.

DR. THRUPP: Other comments there? Yes, Pat?

DR. CHARACHE: Just one. I think this is the kind of question that makes me glad I work for Johns Hopkins and not the FDA. But I think these are the most difficult and also the most critical, and I am relating again to the question of the meningitis, and one of the big ones that a

sgg

lot of people are struggling with is herpes encephalitis diagnosed by spinal fluid positivity and the difficulty of knowing the positive predictive value. The negative predictive value is also difficult to assess in the same setting.

But there is one more factor I would put in here with the example of the tuberculous meningitis. Where one gets the case material now is often in places with very advanced disease. I am thinking of the meningitis building at the Cairo hospital which is full of tuberculous meningitis. But there you run into the problem of a false quantitation issue. You have so many more organisms in your test population when you go to centers like that than you are going to see in this country if you have a child with tuberculous meningitis. So, I do think that these issues where it is so critical for decision-making and for the patient in whom you are making the diagnosis--these are not casual diagnoses--I just do think these have to be handled very carefully and it is just not an easy thing to do.

DR. HAMMERSCHLAG: Thinking of it clinically, because of this issue, if you have a child, a neonate--and, actually, I have a case right now as I am attending a suspected Herpes simplex. Unfortunately, even if you had such a test and it came back negative, because the

sgg

predictive value is 95%, because of the devastating risk of missing, you would still treat because the treatment is really relatively benign. I mean, the risk of treating is far less than the risk of not treating. I think that would also apply to how we deal with TB. So, in situations like that it may have less of an impact clinically. This is something we also dealt with in dealing with the rapid test for group E strep because if you were dealing with that for use in infants for diagnosis of sepsis, you are not really going to rely on that negative test to make your decision whether you are going to treat or not.

DR. NIPPER: I have a question. Theoretically, if you think of Dr. Hadgu's comments that it is possible to do this with mathematics, if you have a good estimate of sensitivity and a good estimate of specificity and you know the prevalence of disease in the population, theoretically you should be able to estimate the positive and negative predictive values reasonably well. I think my two "ifs" or maybe my three "ifs" foul up this thing so badly that you can't do it.

But I am responding to Dr. Charache's eloquent example of a low prevalence disease where you may not have enough clinical material to do a good estimation of a positive and negative predictive value. I don't know

whether any of the people on the panel have experience to say that was either easily done or not easily done. I just don't know.

DR. THRUPP: Well, the bottom line is probably going to come down to the fact that each of these examples, or each target population and each antigen or each test that you are looking at is going to have to be evaluated on an individual case basis, and the guideline is drawn up for each one based on all of these factors. Dr. Edelstein?

DR. EDELSTEIN: I take this question to mean that when there are no other available diagnostic tests, what do you do? In that situation you have to rely on the clinical presentation of the patient, perhaps using a variety of non-specific tests that maybe in combination may yield a diagnosis, or you may need to rely on long-term clinical studies if it is a rare disease, perhaps with follow-up necropsy studies. I don't know. It depends in great part on what the disease is you are trying to diagnose.

DR. THRUPP: Let's move to question three. When can a new device make claims as being superior to the predicate device, the gold standard or the reference method under the following scenarios: the new device compared to predicate; the new device compared to the reference or gold standard; and the new device compared to clinical diagnosis?

Dr. Kroll?

DR. KROLL: I tend to think that the only time the new device can make superior claims is when it is in case c because there you are looking at clinical aspects. You are looking at the entire case and you have a much better idea of truth, and then it is not done in terms of sensitivity and specificity but comparing which does better in terms of assessing, compared to both the previous device and the new device compared to the clinical situation which entails all the information.

DR. THRUPP: I would suggest that this is a circumstance where the data on which this claim is based would have to have statistical evaluation. And if it was a small numbers problem as was discussed earlier, they really shouldn't be able to claim superiority if the numbers were small in the trial or the prevalence was so low that you could never get the numbers out, and these points would have to be made in the package insert. Yes, Pat?

DR. CHARACHE: I think you would also have to be fair to the predicate device and be sure you were studying the same patient population.

DR. THRUPP: As the population upon which the predicate device was first established?

DR. CHARACHE: Right.

DR. THRUPP: Would it be practical to get all that information into the package insert however?

DR. CHARACHE: I think if the company with the new device thought that their device was really superior they would be very happy to do that.

DR. THRUPP: Yes. Dr. Nipper?

DR. NIPPER: What is the difference between a reference method and a gold standard as far as the question writer is concerned? Any?

DR. HANSEN: We have certainly had many discussions among ourselves about that, and you all certainly have mentioned it today. What do we mean when we say reference? Is it an analytical method? Is it a clinical diagnostic method? Micro has a tendency to use a gold standard such as culture. In your world, Henry, you have analytical standards.

DR. NIPPER: Sometimes.

DR. HANSEN: Sometimes.

DR. NIPPER: Yes.

DR. HANSEN: It may be an NCCLS recommended method; it may be a consensus method that we call reference, but there certainly is the separation between analytical, clinical and non-consensus. HPLC may be a reference method for certain things.

DR. NIPPER: I bring that up because Dr. Hammerschlag again put her thumb right on the button about the fact that not all gold standard methods are performed equally well across laboratories. So, my feeling is that a reference method is one which not only carries a certain methodologic principle with it, but it carries a certain imprimatur of minimum precision and accuracy, extremely good technique, well defined according to whatever reagent standards you have and reagent purity--in other words, done right. A gold standard method should then be equivalent to a reference method. I kind of like the idea, in our own minds, of having those achieve equivalence, and demand technical expertise from those who are doing comparative methodologies.

DR. GATES: I just have a quick question in terms of what Dr. Nipper and Dr. Hammerschlag were saying about tests or gold standards varying from hospital to hospital. We are also saying that kind of the sine qua non is comparing it to the clinical diagnosis and, not being a physician, I don't know if it is applicable but do the clinical diagnoses vary from hospital to hospital, or place to place, or doctor to doctor --

[Laughter]

DR. THRUPP: Never, never!

DR. GATES: The other issue though is do I have this straight in terms of what the sense of where we are going here is? I am looking at it from the point of view of industry. Are you saying that if we had a product and we had some predicate product and compared both of those to clinical diagnosis and ours was better than the other one in a controlled study and was statistically valid, we could advertise that ours was the better product?

DR. HANSEN: Don't look at me!

[Laughter]

DR. GATES: Because that is kind of the sense that I am hearing.

DR. THRUPP: If those conditions that you outlined were true and you had the data to support that.

DR. HANSEN: Well-designed studies, well-documented case definitions. That is what we ask from you.

DR. NIPPER: Especially if you publish it in a peer-reviewed journal.

DR. HANSEN: Right. And many of the well-designed studies--products that have received FDA approvals or clearances have appeared in peer-review journals.

DR. HAMMERSCHLAG: What about a lot of things that seem to get approval but don't appear in peer-reviewed

sgg

journals --

DR. HANSEN: Now, now, Maggie!

DR. THRUPP: Let's move on. We want to assign at least a brief period here for any further responses from the audience. Any non-panel member that is in the audience that would like to offer any additional comment or rebuttal? I am not sure if Dr. Green wanted to rebut. Would anybody else like to offer a comment?

MS. POOLE: Anybody from industry who wants to comment?

[No response]

DR. HANSEN: One of the things that Steve and I talked about, and I guess we will have to go through the transcript but could you perhaps give us an overall general summary? I can focus you specifically on what --

DR. THRUPP: With all of the discussion that has gone, it has not exactly been black and white in its conclusions to this point. So, I think what we would like to do is ask the FDA if there are some distillations from this that you would like us to recommend on, hopefully, in a little bit more conclusive fashion?

DR. GUTMAN: Yes, I really apologize. I think we probably should have used more concrete cases. You got diverted by our non-specificity here. But the critical

sgg

issue that is before us, and the best example may be the Chlamydia but it is certainly not a unique or only example--the critical issue is that the technology is pushing at the door, and it is very interesting and exciting technology and, frankly, it has the potential to blow away the predicates, or gold standards, or reference standards or anything else that you want to call them. And we can't wait for mathematical modeling techniques to be devised or become too resource intense in terms of what kind of statistics, but we need some help in communicating to industry, or to ourselves, or to you a way of taking a product that promises to be equivalent or better than what it is being compared with and having some way of defining that and communicating that.

I have no allegiance to discrepancy resolution. So, if the panel as a group thinks discrepancy resolution isn't the right technique, that is fine. I don't even request that you answer this afternoon between now and five o'clock. You can go home and think about it and send letters to Freddie, to Sharon or to me. But the issue is that we have to interact with industry and find out ways, user-friendly ways that we understand, that you understand, and that laboratorians understand and the clinicians understand for appropriately characterizing new technologies

sgg

and communicating that.

That may be a very big order or maybe it is simple, and you said it earlier and I missed it because I was nodding off, but that is what we are looking for, either now or in the weeks that follow. If discrepancy resolution isn't the right tool, then the question is what is the right tool. I mean, one right tool might be to do extensive clinical studies on every new analyte that comes in, and I don't know if that is consistent with the new law that we have been presented with. Maybe it is; maybe there is no way to get around it. But if there is some tool short of a huge prospective or huge clinical study, Sharon and I are all ears. We need to hear about it. Not necessarily now, but we need to hear about it.

DR. THRUPP: Well, one overall quick comment would be that I think we have heard enough comments this afternoon that would suggest that, if feasible, a clinical syndrome, or a clinical diagnosis, or a clinical predictive even if there are no clinical symptoms, like the Chlamydia case, scenario should come close to a gold standard. If the data attest that the new technology comes closer to reflecting that, then it would be valid to shift, or at least to modify the so-called gold standard at least under defined circumstances. That is kind of an overall comment. I think

sgg

everybody has been after wanting clinical validation when that is feasible. Dr. Charache?

DR. CHARACHE: Just two other thoughts, we have commented that not every laboratory is going to be as helpful or as informative. We know that the laboratories with the highest percent positive rate by culture show the least advantage of the non-culture techniques. But it is also the group that is the easiest to interpret.

Just looking at your Table 4 for the Chlamydia study, sensitivity between the labs, as I read it, varies from 92.9% sensitive to 53.8% sensitive. I think a lot can be done by the company that is setting up the assay to avoid discrepancies through ensuring consistency of approach. If everyone says they are using the CDC standard method for tuberculosis, they should really be using it, and not all over the map which we saw they were doing when those studies were evaluated. So, I think that is one thing. I think that there is a lot that can be done to select the laboratories that are going to minimize this problem.

I think, secondly, if you are going to have a resolution strategy, for many diseases clinical is going to be the best you have. It won't be the only one. For Chlamydia it is a flawed one. One thought that I would stress is that under no circumstances would I use a test

sgg

that was in itself experimental, or home brew, or not thoroughly reviewed as a means of validating one that you are applying to assess. You don't want to use a different targeted series of primers without knowing, for example, if those primers might even cross with another organism, or something else. I mean, I don't think you want to validate a test you are trying to define by a non-validated approach. So, I think there are things that can be done to simplify, and I am a believer of simplifying all of these things just as much as you possibly can.

DR. THRUPP: Pat, how would you respond to the argument that, let's say, in a trial, in a developmental trial that a manufacturer did select a range of laboratories and the patient populations were carefully defined and any additional drugs that patients were taking or medications that might have represented inhibitors or a number of factors in terms of the possible technical variabilities in the LCR, or whatever the molecular test, were reasonably well controlled for and, yet, in certain laboratories, without apparent reason, their culture "gold standard" technology was deficient and they had a very low sensitivity, and not explained directly? That could be used as an argument to say that in the real world culture techniques are in a certain percentage of laboratories going

sgg

to be poor, therefore, the molecular method should become the method of choice. How would you respond to that argument? Is that a valid argument that might well be seen in the marketplace?

DR. CHARACHE: I think the argument--I mean, I have seen this in a lot of models. The fluorescent microscopy detection of respiratory syncytial virus--in a good lab you always get a higher return by culture. In the real world you usually get a higher return by fluorescent microscopy because of the problems that are associated with transport, and what-have-you, for a very fragile agent. But I think there is a lot that the company that wants to present this test can do. I am not saying that perhaps in the real world it can be wiser to use a non-culture technique because I believe that can be the case, but we are not talking about that here. We are talking about validating the fact that this is a safe and wise thing to do.

If you look at five chlamydial laboratories, we blind passage everything at 48 hours. Many labs don't. It depends on what the McCoy strain is that you are using. There is a tremendous amount that can be done by anyone who wants to get into this to make their lives a lot easier if they have somebody who knows the microbiology and what they

sgg

are looking for and pre-review the labs they want to use, and perhaps make sure that you ship them all the cell line you want them to use, or whatever you are doing. So, I think that there is a great deal that can be done to assist the companies that are not based on microbiologic background to do these.

DR. THRUPP: That is true in terms of the validation in trials that are presented --

DR. CHARACHE: Right.

DR. THRUPP: -- but the package insert that the FDA has to also work with the company to produce has to be realistic, and would have to address the applicability in the field issues also.

DR. CHARACHE: No, because what the FDA is, I am sure, trying to do in comparing these methods--the target is not how well does one method compare to another. That is just a strategy to get at the real target, which is how sure you are if you get a yes with this test that the patient has this infectious agent or this analyte, whatever it is.

DR. THRUPP: Yes. Dr. Ogandi, you had a comment.

DR. OGANDI: Yes. I need to agree with the concept of being careful to select laboratories. If you select a laboratory that has just done one of those tests in a couple of months and you select another laboratory that

sgg

has a very high volume and does these on a regular basis, you will have a real difference in what you get out of that. So, the industry should do a little more in choosing laboratories that are involved in this and have the expertise so that you can have something to compare with.

Also, when we talk about the culture methodology, it seems as if the advance in technology is not in that area. But it is also in that area because what cultures used to be, many of those are changed. So, there are advances in all the areas, but I think selecting laboratories and the expertise before you do these to compare--if I am doing DNA I could select some laboratories that wouldn't know where to start and you need to use it to compare results so that you could sell something. So, I think selecting laboratories will be an emphasis.

DR. THRUPP: That is true, although I am not sure that the FDA has the resources or the authority to direct the manufacturer how to select their test sites. Am I right?

DR. GUTMAN: That is absolutely right. That is why we have this emphasis on truth in labeling so that we do the best we can and negotiate with companies. Then, whether they have a gold submission, a silver submission or an alloy submission, we like to try and communicate it in the

labeling.

DR. CHARACHE: I think I was mostly speaking to the companies because there is nothing that makes me feel more depressed than seeing a couple million dollars worth of work in which they just didn't know the microbiology or the chemistry or the hematology, whatever it is, so that they have wasted it because they haven't set it up well. So, I am looking forward to your being able to help them.

DR. THRUPP: Dr. Gutman or Dr. Hansen, in terms of responding to specific concerns of the FDA that we have only given you waffling responses to, is there something else that you would like us?

DR. GUTMAN: Well, I have to make an observation because it has been a very interesting day for me personally, but it is so interesting because we had some preconceived notions, for example, about the first case that we presented, and there were a number of people who cooked up these cases and we thought surely there was nothing else that the panel was going to say. It was going to say that when you got to 1c you had a very reasonable study and, Henry's aspersions aside, you would be able to clearly say this is a study where you could say clinical sensitivity and specificity are characterized and a predictive value is a reasonable thing to put into the labeling. You didn't say

sgg

that. I am not sure we are going to go away and discard that as a practice but I was personally surprised you didn't gravitate towards 1c.

Then I was more amazed--we had lunch as we came back and we were talking about a really difficult situation in which we were taking bank samples because we had this rare disease that was 1/80,000, and I thought I was hearing you say that in that case, even though they were obviously very carefully selected samples, I thought I actually heard some enthusiasm for using sensitivity and specificity.

I guess these are just really treacherous and tricky issues and there are a lot of semantic problems here, and I view this as a starting point to maybe dialogue with you folks as we try to develop some guidance. We do hope to develop some guidance. We do hope to work with industry, and we do hope to find some solutions and not maybe always just present you with questions. But we have a long way to go because we certainly perhaps had too many questions not phrased as well and I don't know that--I know there are a lot of smart people here and I don't know that we have particularly fabulous resolution and it may involve the difficulty of issues, not the quality of any of the people framing the questions or answering them. I hope that is the case.

DR. THRUPP: You brought up 1c. I am not sure that the discussion was that negative about 1c. I mean, this was the ELISA type device, but the 1c example was where the population targeted was a known population and you had a black and white disease-based gold standard, if you will, and I thought the discussion indicated that in that scenario the conclusions were reasonable, if I interpreted that correctly. So, I think the bigger problems were in the 1a and 1b type examples. Sharon?

DR. HANSEN: One of the things that I would like to add to what Steve is saying is, again, with the new law, in the Class III area with PMAs or PDPs we are encouraging, and the companies can come to us and they are being encouraged to come before they start the studies so we can help establish protocols. But we would like to extend that really to all the high risk devices so there will be a learning curve. Certainly, I would think that those of us that are going to be involved in developing outlines for clinical studies and things like that in the laboratory world would be able to call on you for advice and consultation, as well as the industry. The industry knows who the panel members are and I hope you don't turn them away if they ask for help because our purpose is to try to help the industry, to try to have good products in the

marketplace.

DR. THRUPP: There are several questions and I was perhaps skipping some of the specific questions in the interest of time. We are actually finishing up ahead of schedule, which is unheard of I guess. But are there any of the specific questions that we kind of skipped over? This is one that I skipped because we talked about it a lot before, but Dr. Nipper points to 3, 1c.

DR. NIPPER: I was interested because I wanted to learn a little bit about how you all thought Chlamydia discrepant results should be resolved. I am learning about the clinical picture of Chlamydia today and so I am just curious about what that answer would be in this particular case.

DR. THRUPP: How and when should discrepant results be resolved? That was discussed a lot.

DR. HAMMERSCHLAG: Actually, Chlamydia is really easy when you think about it. Wait until we start getting into Chlamydia pneumoniae one of these days and then we are really going to have fun.

I was sort of coming to the conclusion, personally, that rather than discrepant we should be having studies that would run the nucleic acid amplification test in culture with the new test in parallel.

DR. THRUPP: With the arbitrating test --

DR. HAMMERSCHLAG: The arbitrating test.

DR. THRUPP: The alternate test in parallel. That obviously adds to costs of the trials, and that might be an example where it is more feasible. Let's suppose, for purposes of discussion, that that process was not feasible because it was either too difficult or too expensive, or whatever. Then we would be back to the new versus the predicate and then what to do with the discrepancies, and we have heard much discussion about the biases that the selected discrepant case analysis leads to.

On the other hand, we have also heard, from Dr. Green's review, that the bias can be calculated and that in many scenarios it is small. But I think I got the sense that perhaps the majority of discussants are at least cautious or skeptical about the selective retesting of only discrepant results and having that be a standard procedure under most circumstances. Dr. Todd?

DR. TODD: I think that it is important to define at the beginning what you are going to use as a positive. So, that would be any kind of clinical syndrome, any kind of clinical diagnosis that you can include. Then what has been the gold standard, the culture technique, and then if there are any other tests that you want to run should be run on

every sample initially, along with the new test. If the new test is proven to be the new gold standard, it is actually going to be a decrease in the final cost.

DR. HAMMERSCHLAG: Except that for Chlamydia you really can't apply a clinical diagnosis, and most of these tests are being used, again, in a screening situation in frequently asymptomatic individuals.

DR. PEPE: I wanted to ask, Dr. Hammerschlag, supposing in the ideal case where you could get the three different test results that you talked about, what would you use then as the comparison for the new test? Would you use a culture positive or a confirmatory test that is positive?

DR. HAMMERSCHLAG: I think you could probably use that. If you ran them in parallel that would sort of obviate the situation of having to do selective discrepant analysis. I have to say I have done discrepant analysis myself, but in a smaller population dealing with a clinical situation. Basically I have been dealing with Chlamydia ophthalmia but the presentations and the background papers have left me feeling a little uncomfortable. You know, I didn't realize there was a potential of this problem. I can see in some ways where it is coming from. Certainly the issue of an in-house test, like the MOMP assays that both LCR and the AmpliCor have which, by the way, have not been

sgg

independently evaluated and one would probably assume would not be as sensitive as the plasmid-based assays because, you know, there are 10 copies per plasmic per Chlamydia cell but only 1 copy of the OMP-1 gene. So, that is an issue, plus the fact that you may perpetuate some of the same errors with the same technology. But now we have an opportunity where we have some variety and we can pick. In the end it may end up actually being less work because it is all run together, and you don't have to go back and retest and play around with some of the specimens.

DR. EDELSTEIN: I have heard a lot today about the theory and potential bias of discrepant analysis and, in fact, have heard two analyses that were discordant themselves --

[Laughter]

-- I wonder whether it would be reasonable to sponsor or encourage some actual proof of principle, in that various statistical approaches to resolving these discrepancies be formally studied with actual clinical specimens. For example, someone could do random sampling of each cell. Another possibility is to test all the specimens to see what the incremental yield would be in terms of reducing bias and see actually in which direction the bias goes.

DR. THRUPP: I think that to some extent some of that has been done in some of the papers that were presented.

DR. EDELSTEIN: Well, those are all theoretical though.

DR. THRUPP: Retrospective analysis --

DR. EDELSTEIN: Yes, but none of them, as far as I know, actually involved retesting all the samples tested.

DR. THRUPP: Was there another hand up? I think Dr. Gates was next and then Dr. Nipper.

DR. GATES: I guess looking at it from the other perspective, I think if industry is required to test against something, have a predicate, and we have all agreed that in some cases the technology has outstripped the gold standard, so unless we do come up with some discrepant analysis in some way we are always running a risk that we are putting a damper on any new technology because all we are testing against is stuff that we already know doesn't work as well as it ought to.

I kind of go along with what Dr. Hammerschlag was saying. I am thinking, well, what do you do if you have culture positive and nucleic acid negative and test positive, or something like that, and you are back in the same box. But the other thing is what Dr. Edelstein was

sgg

saying, that there be some sort of standardization or something like that where everybody could agree that if you met whatever that standard is, that is all you need to do. I think that may be a good direction to go in.

DR. NIPPER: The pebble I have in my shoe about this issue is on this slide.

[Slide]

Maybe I have missed something and I don't know enough about micro to know what I am talking about here, but the thing that bothers me about this slide, and I think it is the dilemma that Sharon and Steve were talking about, is what do you do about those 14 patients up there who are symptomatic, who have 3 of the tests that are negative and a nucleic acid test positive, and you are labeled false positives? That bothers me because in my ignorance about this particular situation I wonder if the NAA test is telling the truth and these people, even the 11 that are asymptomatic within the endocervical samples and the 7 urines that are asymptomatic, if those people need treatment and they are not going to get it because we, somehow or other, mislabeled--I shouldn't say mislabeled; that is a regulatory word--if we called this something that it is not scientifically, and if we are also missing the boat because we don't go back and try to figure out with the discrepant

analysis what is going on with those patients. That is what is bothering me about this particular issue.

DR. THRUPP: Let me give you a quick response on that very scenario. You picked out those 14 patients that have symptoms and have a positive NAA but all the other tests that were run are negative.

DR. NIPPER: Right.

DR. THRUPP: It is entirely conceivable, and in some other tests there may be examples where these individuals happen to be colonized with a Bacteroides or with a peculiar Proteus or something that has a little bit of nucleic acid that cross-reacts with this assay. You can't necessarily jump to the conclusions that these symptoms, whatever they are, are related to Chlamydia if you are testing a new test, where all these others are negative. So, that could well be a false positive.

DR. NIPPER: The problem is should we go back and test 2,900 samples in order to resolve that issue or should we find out what is going on with those specimens?

DR. HAMMERSCHLAG: Number one, I agree with Dr. Thrupp. You can't make the assumption that those symptoms are due to Chlamydia. As a matter of fact, the predictive value, especially in women, of clinical symptoms is terrible for predicting who is going to be infected and who isn't

going to be infected. What is going on? I don't know. I mean, is that the purpose of thing to find out that there are false positives? Again, it comes down to how fine do we have to sharpen the point of the pencil? We are never going to resolve it completely--ever, ever in this space-time continuum. I think we have to realize that we are going to maybe approach perfection but we are never going to achieve it, and we have to determine what our minimum requirements are and to educate people as to the limitations of this test. Number one, you will always probably have a false positive here and there or a false negative here and there.

DR. NIPPER: So should we leave it alone?

DR. HAMMERSCHLAG: I don't know how much we can sharpen this pencil.

DR. CHARACHE: I was just going to make two points. One is that the symptomatic women may very well have GC or some other disease. So, I think we would be doing them a greater disservice by always assuming that if they are symptomatic it must be Chlamydia than if we assume that we don't know what it is and we are going to treat them accordingly.

DR. THRUPP: Particularly for evaluating a new test.

DR. CHARACHE: Yes. I think that for many tests,

sgg

as we have discussed, you can get a good clinical correlation, and I think that should be. Just like your system 1c. I think that is our first fall-back, and I think then you can get your predictive values, and I think you can define your predictive values in terms of the populations in which you want to use the test.

DR. GUTMAN: But Henry asked the right question, and it is a tremendous burden, and manufacturers want to hear the answer. Do you then go back and characterize all 1,300 or do you characterize the 19 oddballs?

DR. CHARACHE: I think if you want to characterize anything further or go back into that same specimen again--I personally would think that you would want to do neither; you wouldn't want to do only those 19 but you would want to do an appropriately selected subset of the others. I think you have to do that. A number of these newer assays are using the biotin markers and protease has biotin. I mean, there are lots of bacteria that can cause the same error. So, I think we just have to be very circumspect.

DR. GUTMAN: Although a subset analysis sounds better than 1,300 probably to the manufacturer.

DR. CHARACHE: Well, personally I would never suggest all 1,300.

DR. THRUPP: Unless you were dealing with a very

sgg

low prevalence problem and you really wanted to pin it down.

DR. CHARACHE: Even then I wouldn't.

DR. THRUPP: Do we have any other suggestions or comments? If not, I would like to thank all the presenters, all of the audience for their patience and all of their panel members for a very interesting session. See you tomorrow.

[Whereupon, at 5:25 p.m., the proceedings were recessed, to be resumed at 9:30 a.m., Thursday, February 12, 1998.]

- - -