FOOD AND DRUG ADMINISTRATION

CENTER FOR DRUG EVALUATION AND RESEARCH

MEETING OF THE

DERMATOLOGIC AND OPHTHALMIC DRUGS ADVISORY COMMITTEE

8:10 a.m.

Tuesday, November 5, 2002

Versailles Ballroom
Holiday Inn - Bethesda
8120 Wisconsin Avenue
Bethesda, Maryland

ATTENDEES

COMMITTEE MEMBERS:

ROBERT S. STERN, M.D., Chairman
Professor of Dermatology
Beth Israel Deaconess Medical Center
330 Brookline Avenue
Boston, Massachusetts 02215

KAREN M. TEMPLETON-SOMERS, Ph.D.
Acting Executive Secretary
Advisors and Consultants Staff
Center for Drug Evaluation and Research
Food and Drug Administration
5630 Fishers Lane, HFD-21
Rockville, Maryland 20857

ROBERT KATZ, M.D.
Dermatology
11510 Old Georgetown Road
Rockville, Maryland 20852

LLOYD E. KING, JR., M.D., Ph.D.
Professor of Medicine, Dermatology Division
Vanderbilt University
1301 22nd Avenue North
3900 TVC (The Vanderbilt Clinic)
Nashville, Tennessee 37232-5227

PAULA KNUDSON, Consumer Representative
Executive Coordinator for IRB Research and
Academic Affairs
The University of Texas
Houston Health Science Center
1133 John Freeman Boulevard
Jesse Johnes Library, Room 322
Houston, Texas 77030

R. TODD PLOTT, M.D.
Acting Industry Representative (non-voting)
Vice President, Clinical Research
Medicis Pharmaceutical Company
8125 N. Hayden Road
Scottsdale, Arizona 85258

ATTENDEES (Continued)

COMMITTEE MEMBERS: (Continued)

SHARON S. RAIMER, M.D.
Professor and Chairman
Department of Dermatology
University of Texas Medical Branch
Galveston, Texas 77550-0783

KATHLEEN Y. SAWADA, M.D.
Dermatologist
U.S. Naval Reserve Medical Corps
Alpine Dermatology Associates P-LLC
1785 Kipling Street
Lakewood, Colorado 80215

MING T. TAN, Ph.D.
Professor of Biostatistics
Room N9E36
University of Maryland Greenbaum Cancer Center
22 S. Greene Street
Baltimore, Maryland 21201

CONSULTANTS: (voting)

S. JAMES KILPATRICK, JR., Ph.D.
Professor of Biostatistics
Medical College of Virginia
Virginia Commonwealth University
1101 East Marshall Street
Sanger Hall, Room B-1-039-A
Richmond, Virginia 23298-0032

THOMAS R. TEN HAVE, Ph.D.
Department of Biostatistics and Clinical Epidemiology
University of Pennsylvania
School of Medicine
607 Blockley Hall
423 Guardian Drive
Philadelphia, Pennsylvania 19104-6021

ATTENDEES (Continued)

FOOD AND DRUG ADMINISTRATION STAFF:

MOHAMED ALOSH, Ph.D.
JONCA BULL, M.D.
JONATHAN WILKIN, M.D.

ALSO PRESENT:

DONALD A. BERRY, Ph.D.

C O N T E N T S

P R O C E E D I N G S

(8:10 a.m.)

DR. STERN: I hope everyone is refreshed. I at least found it an interesting day yesterday and helpful in orienting myself to the questions at hand. So I'd now like to call to order the second day of the meeting of the Dermatologic and Ophthalmic Drugs Advisory Committee to the Food and Drug Administration.

This morning we'll first have an open public hearing. I'm sorry. We'll first go around the table again, starting with Dr. Plott.

(Laughter.)

DR. STERN: We'll start on this side.

DR. BULL: Good morning. Jonca Bull from the Office of Drug Evaluation V.

I would like to extend our thanks to the committee for such an invigorating discussion yesterday, taking time from your busy schedules, and we look forward to hearing your input on the questions today. But thank you so much.

DR. WILKIN: Jonathan Wilkin, Division of Dermatologic and Dental Drug Products, FDA.

DR. KATZ: Robert Katz, a dermatologist in practice, Rockville, Maryland.

DR. RAIMER: Sharon Raimer, Professor of

Dermatology at the University of Texas in Galveston, a committee member.

DR. TAN: Ming Tan, University of Maryland School of Medicine, preventive medicine, epidemiology.

DR. TEMPLETON-SOMERS: Karen Templeton-Somers, acting Executive Secretary to the committee, FDA.

DR. STERN: Robert Stern from Boston.

DR. SAWADA: Kathy Sawada, practicing dermatologist, Denver, Colorado.

MS. KNUDSON: Paula Knudson, IRB administrator, University of Texas, Houston.

DR. KING: Lloyd King, Vanderbilt Dermatology, Nashville VA, a member of the committee.

DR. TEN HAVE: Tom Ten Have, biostatistics and epidemiology at the University of Pennsylvania School of Medicine.

DR. PLOTT: Todd Plott, Vice President of Clinical and Regulatory, Medicis Pharmaceutical. I'm the acting Industry Representative.

DR. STERN: There are some conflict of interest statements to be read.

DR. TEMPLETON-SOMERS: The following announcement addresses the issue of conflict of interest with respect to this meeting and is made a part of the record to preclude even the appearance of such at this

meeting.

Since the topics to be discussed will not have a unique impact on any particular product or firm, but rather may have widespread implications with respect to an entire class of products, all committee participants have been screened for interests in products indicated for use in the treatment of acne vulgaris and their sponsors.

In accordance with 18 U.S.C. 208(b)(3), Dr. Thomas Ten Have and Dr. Robert Stern have been granted particular matter of general applicability waivers which permit them to participate fully in the matters at issue.

A copy of the waiver statements may be obtained by submitting a written request to the agency's Freedom of Information Office, room 12A-30 of the Parklawn Building.

Because general topics impact so many institutions, it is not prudent to recite all potential conflicts of interest as they apply to each member and consultant.

FDA acknowledges that there may be potential conflicts of interest, but because of the general nature of the discussion before the committee, these potential conflicts are mitigated.

Lastly, we would like to note for the record that Dr. R. Todd Plott is participating in this meeting as a non-voting acting industry representative, employed by

Medicis Pharmaceutical Company. Medicis Pharmaceutical is one of the many firms which could be impacted by the committee's discussions.

In the event that the discussions involve any other products or firms not already on the agenda for which FDA participants have a financial interest, the participants' involvement and their exclusion will be noted for the record.

With respect to all other participants, we ask in the interest of fairness that they address any current or previous financial involvement with any firm whose product they may wish to comment upon.

Thank you.

DR. STERN: Thank you very much.

We now enter the open public hearing, and we have one person who has indicated to us so far -- and we'll ask for others -- who would like to speak. This is Dr. Donald Berry of Berry Consultants. Perhaps he could start off by telling us what Berry Consultants consults about.

DR. BERRY: Thank you very much, Dr. Stern.

My day job is at the University of Texas, M.D. Anderson Cancer Center, where I'm chair of biostatistics. I have with my son a consulting company, and we consult with pharmaceutical companies, including Stiefel. I'm a paid consultant for Stiefel.

I want to do two things today. One is to talk about outliers in the context of analyzing acne lesion counts. I'll give you an example of the effect of outliers and give you a real data example and address the question of taking logarithms.

And the second part of my presentation is addressing what is an important question not only in dermatology but in cardiology, oncology, and essentially every medical application. There are sometimes a number of studies that address the same question, and at issue is how to combine. It is a mistake to simply regard the many studies as a single study and throw everything together, and the question is how to do it in a reasonable way. That's the second part of my presentation.

This is two data sets that I generated artificially. Think of it in terms of a numerical reduction in lesion counts, but what they are in fact are random samples from a normal distribution in the case of the first one, treatment A, which has mean 0. So nothing is happening here. And in treatment B, there is a 10-count difference. So there's a shift in the two things.

These are bell curves. These are normal distributions that are shifted by 10 points. The standard deviation in both cases is 20. I simply generated 25 observations from each, did a t-test and found that I had a

statistical difference between the two at p equals .013.

Now, what I want to do to show you the effect of outliers -- and I understand that you talked about this to an extent yesterday. Suppose I take any one of these observations and change it. And I chose the one which is the worst one for drug B. It's the one which had a minus 17 in terms of differences. So if this were to change, it would have an impact on what you conclude.

I could take any other one. If I took something on this side, I would move it down to show the same effect that I'll show by taking this minus 17 and moving it up. So it's going to become better for drug B. What you expect is that this p value is going to get smaller because the inferential impact is favorable for drug B.

So here is the significance level for A versus B. This is the p value. This is for that minus 17 observation being down in here. So this is minus 17. There's the .013 p value associated with minus 17. And now I moved it up. I moved it up to bigger than 0 up to 60 and 80. And what happens is exactly what you would expect to happen; namely, the p value gets smaller. The inference is stronger that there is a benefit in favor of B.

But then a surprising thing happens. It starts to get worse. It doesn't make sense because what is

happening is the drug is looking better. There is a bigger reduction in lesion count, as I move this number up. Here you see enormous reduction, up to 200. But now it's even worse at 200 than it was at minus 17. And what happens if I keep going -- and these extreme values are something that you observe frequently with a percent reduction in lesion counts -- is I lose significance completely, and indeed, the asymptote, if I were to keep going on this, is p value equals .32, not even close to being significant.

So take any data set, modify one of the observations, and you'll get essentially .32 as the two-sided p value if you move it far enough. It doesn't make sense. It means that outliers can completely ruin any inference.

What's happening is the t-test. The t-test is the ratio of the mean to the standard deviation and what's happening here is that both of them are becoming big and the ratio is becoming essentially 1 standardized.

That happens the same for any value in any direction.

It's even more problematic if there are multiple outliers, although I should draw a line there. If half of the data are outliers, then there are no more outliers. So it's multiple outliers but for a small proportion.

It's problematic for skewed distributions if there's lack of normality in having substantial skewness or fat tails, so to speak, observations that are, in effect, outliers.

It happens if you do percent change from baseline. The problem is that if you start with a modest baseline value and there is a substantial increase, your inferences go down the tubes.

So you've got to do something, and what should you do?

One possibility, the simplest possibility, is to truncate. It's an age-old statistical technique. You simply cut the value at some arbitrarily chosen point and you say I'm not going to count it if it's bigger than that. I'll round it down to that point. So percent reduction. If it turned out to be minus 700, I would round it down to maybe minus 100.

Another possibility is to rank the counts. Ranking would make that observation at minus 17, when it got bigger than the biggest observation in the data set, it would have the same rank throughout, and so your conclusions wouldn't change.

Another possibility, a nice possibility I think because it maintains the clinician's understanding of percent change with a mild modification, is to take -- this

is the usual percent change -- take the post-treatment value, subtract the baseline, and divide by the baseline. The symmetrized percent change is to look at the -- symmetrized in the sense of time with respect to the post-treatment and baseline, essentially considering the change as it depends on the mean of the two observations, the post and the baseline.

Still another possibility is to take the logarithms of the counts. This is a standard thing in scientific inference. Roughly half the time you should take logarithms, do the analysis in logarithms, and then anti-log back so you understand what the original scale was. This half -- I'm just picking a number. A substantial proportion of the time, logarithms is the appropriate analysis point.

When you take logs, for example, if you compare the post-treatment with the baseline, you might take the logarithm of the post-treatment and the logarithm of the baseline and subtract the two. In a way it's like proportion change because it's the logarithm of the ratio of the post to the baseline. But this might overcompensate. You're taking something which is skewed and potentially skewing it in the other direction. I'll give you an example of that.

There may be 0's in the data. How do you

handle that?

A possibility, a standard thing to do when there are 0's, is to add something to get rid of the 0's, to add a constant C. Almost always C is 1. That's because 1 is the first number. It really matters what you add. It matters very much. If the units are counts and you add 1, that's a very different effect than if the counts are, let's say, in 10's and you add 1.

So what's C? I want to address what's C.

This is taking a particular actual data set of two treatments, B and X. This is a histogram of the distribution of the difference between the logs post and pre for a particular value of C. C equals .5 which is essentially the same if C were equal to 1. And you see what happens is that there is substantial skewness in the opposite direction from what we're correcting for. That is, there are large observations out here that are going to affect things. I don't know if you can see that, but there are some observations in this tail of the distribution.

The p value for the comparison of X versus B. These are B's and these are X's. They both have roughly the same distribution, and I've just combined them here to accentuate. The p value for the difference between the effect of B and the effect of X is .09. So it's not statistically significant.

This is the same example with a different C. Now I chose C equals 29. I'll say why that in a minute. But what you see is essentially a bell shaped curve. Indeed, C was chosen to make the curve bell shaped. This is an age-old technique in statistics, to transform the data so that you get normality. It's somewhat less important in the modern era than in the previous era because of the possibility of doing computations now that were not available in the 19th century, say.

Now when you transform to normality with this value of C, the p value for the difference between the two treatments is statistically significant. P equals .03.

In transforming the data toward normality, a possible thing to do to make it look more normal is to minimize the sum of skewness and kurtosis. Skewness is the third moment. It's a measure of how skewed the distribution is. Kurtosis is the fourth moment. It's the fatness of the tails. So it measures outliers, although skewness takes into consideration outliers as well.

Minimizing skewness and kurtosis preserves the false positive rate and it preserves the power. This is a consequence that's shown in this paper.

So here's what happens in this actual data set. Remember I did C equals small, C equals .5, and I got a p value of .09, not statistically significant. As you

increase the value of C -- follow the green curve here -- the skewness plus kurtosis gets lower. It's becoming more and more normal. It hits the maximum fit to normality at C equals 29, and then it goes off in the other direction. So what this technique says to do is to pick the one that's most nearly normal. You preserve the false positive rate. You preserve power, and you get an analysis which is more finely tuned to the normal distribution. And for there, the p value is, reading this off in this direction, .03. So you're taking something which is not significant. It becomes significant, but it's right because you're transforming to normality and your technique is based on the assumption of normality.

Some important issues, to summarize this part of the presentation. It's essential to specifically the analysis that you're going to do in the protocol, and don't be ambiguous about it. Don't say, well, if this happens, we're going to do that unless the "this" is very specifically described. Best to say here is the method. It's unambiguous.

The method should be robust and among the robust techniques are truncation. This is really crude. You should never use change from baseline without specifying one of these: truncation using ranks, a symmetrized percent change, or doing this log of the count

plus C.

So that's the first part. I want to talk about the second part, combining study results.

There are two levels of experimental units when you've got several studies to be analyzed. One is within each one of the studies, you have patients. But the study is an experimental unit. That's where the hierarchy comes from. It's the hierarchy of experimental units. And you could go further to have three or four types of experimental units.

You can't simply combine patients from different studies for all kinds of reasons. They tend to be different because of the different geographical area. They tend to be different even if the study protocol is the same. They tend to be different because the clinicians involved admit patients of a different type within that predescribed eligibility criteria. So it's a mistake to simply throw them together. Everybody knows that I think.

The inferences that one makes in doing a hierarchical analysis is to conclude something about the population of studies and the treatment effect within that population of studies, but also you can ask what is the effect in my particular study. So there may be a reference study and you're borrowing from the other studies in this hierarchical fashion.

So here is a generic example that has nothing to do with lesion counts. It's a single treatment, just to show you what happens here, show you the effect. These were nine studies with 20 observations in the first study. This is 20 patients and there were 20 successes. This is actually depression. There were 20 successes among 20 patients who were depressed in this study. In this study there were 10, and there were 4 successes. In this study 16, and there were 11 successes, et cetera. A total of 150. 106 of them were successes.

So is the ratio 106 to 150 the appropriate estimate for the benefit of this treatment? More interestingly is the precision associated with this, the precision that you would get from 150 patients in a single study. And the answer is no.

So this is a picture of those points. This was the 20 out of 20. This was the 4 out of 10. And I've shown the dots here on this scale, the area roughly in proportion to the sample size in the particular studies.

The pooled analysis, simply ignoring the difference in studies and throwing everything together, gives this likelihood function that you see here. This is an estimate of this r effect, the success rate. So it's looking pretty tight. A hierarchical analysis, recognizing the possibility of heterogeneity in the studies and doing

this experimental unit stuff, which is the study is an experimental unit in itself, gives much less precision associated with your conclusions about the success rate, r.

Using a Bayesian analysis, a Bayesian modeling in which you borrow from the other studies in the hierarchical modeling, you shrink toward the overall means. So these were the original pictures. These are the dots unadjusted, and these are the dots adjusted for this borrowing, viewing study as having itself a distribution. So a particular patient in study 1 contributes to the conclusion about study 1 but also to the conclusion about study 2, much less about the conclusion of study 2 because, of course, it wasn't in study 2, but through this mechanism of the study having a distribution, the effect of study 1 plays a role in study 2 because they were both from the same distribution. And the Bayesian analysis borrows more if the data are comparable and less if they're not. Indeed, if they're very dissimilar, then there's essentially no borrowing.

Not this. It's a mistake scientifically to suppose that the r in each one of these studies is the same.

This is an example, a clindoxyl example, in which there are five studies. These are the percent reduction in total lesion counts. I'm doing what I said I

shouldn't do, percent reduction. The reason is that this is what was in the protocol. So you see study 1 had four treatments. Study 4 had only three treatments. There was no vehicle in study 4. So these are the mean percent reductions by study and these are the sample sizes. So in study 1, the vehicle had a 1 percent reduction; benzoyl peroxide, 20 percent in study 1; clindamycin, 11 percent; clindoxyl, 41 percent.

Putting that table into a picture, you see these are the five studies, again roughly in proportion to their sample size for the comparison of clindoxyl and benzoyl peroxide. This diagonal line is where they're equal. So what you see is in all five of these studies, the clindoxyl had a lower reduction than did BP.

There ought to be some way to put those together. Simply throwing them together and saying that there was one study is wrong. But recognizing that they're looking at the same question in roughly the same population, although heterogeneity is certainly possible in this type of analysis -- and indeed, it's exquisitely tuned to detect heterogeneity. There ought to be some way to put these together into an overall conclusion.

These are the Bayes estimates by studies, these red dots. This one, number 4, is pulled into the mean. So there's regression going on in this direction, regressing

to the mean in the BP direction, regression to the mean in the clindoxyl direction, and regressing to the comparability of the two.

This is the conclusion. These are the probabilities that each treatment is better than clindoxyl by study and overall. So, for example, in study 1, vehicle gel -- remember that was 1 versus 41 for clindoxyl. This probability that a treatment is better is a Bayesian conclusion not a p value. The probability that the vehicle is better than clindoxyl is essentially 0, that BP is better than clindoxyl is, again borrowing from the other studies, .8 percent. Clindamycin is .1 percent.

There's an overall reading here too. This talks about the population of studies. Think of a new study coming from the population.

Study 4. I put this thing in red just to highlight. You remember that study 4 had no patients assigned to vehicle. We can still, in the context of study 4, ask the question how would it be in comparison to vehicle. We couldn't do that without the other studies. The other studies allow us to compare clindoxyl with vehicle, and so there's a borrowing that goes into study 4. This borrowing in the overall -- we, of course, don't know what the full population of studies looks like. We have only five, and within each five, we don't have a full

certainty about what the effect is in that study. But it's these two types of variability that are going together here. But still, in this example, it spells pretty clear comparison of the two.

So what did I do? I talked about the resolving outliers and resolving skewed distributions, gave you an example of the effect of outliers, discussed taking logarithms and how you can transform to normality and get a stronger conclusion. The second part was how to combine study results using hierarchical modeling in a particular example.

So that, Dr. Stern, completes my presentation.

DR. STERN: Thank you very much.

Are there any questions from the committee for the presenter? Thank you. Oh, I'm sorry.

DR. TAN: Professor Berry, I just wanted to ask in a hierarchical model combining the studies, what is the hyperparameter specification for the prior in a hierarchical model?

DR. BERRY: Let me tell you about the first example that I gave. What we did is suppose that there is a distribution of the proportion. The distribution of the proportion was beta. I apologize to people who don't know what that means. It's a beta distribution. I know you know what that means, Dr. Tan. It's the parameters of the

beta, of which there are two, are the hyperparameters and we put a probability distribution on that. And the distribution that we put on that was noninformative. So what that means is that the data are telling you what that distribution is and it's not telling you a particular beta distribution. It's giving you a probability distribution over that hyperparameter set. Remember that picture with the pooled estimate and then the hierarchical analysis. The hierarchical one, the one that was very flat, that was the average of those beta distributions with respect to the posterior distribution of the parameters.

DR. TAN: Okay.

DR. STERN: Dr. Alosh would like to ask a question, if it's okay with you.

DR. BERRY: Sure.

DR. STERN: We need your permission for an FDA person to do it. That's fine?

DR. BERRY: Oh, absolutely. I love the FDA people.

(Laughter.)

DR. ALOSH: Okay, thank you. Thank you for the illustration about taking one observation outlier on one side and seeing how the p value changed as one makes that outlier extreme.

My question really has two parts. The first

part about the log transformation and adding constant. One point, as you know, we would like things to be prespecified in the protocol, which you touched on this by saying the analysis plan should prespecify the stat analysis.

However, I don't see how choosing a constant will fit into that because it seems to me adding a constant to the data, you cannot prespecify it because this is data driven in a way. And someone could change the constant until probably you reached that significant p value.

There is an issue also about interpreting the data after you add a constant. Personally I don't know whether you were here yesterday or not, but I think I'd prefer the rank transformation, and we have other statisticians here who might jump in. So I'd like to see the comparison. What's your comment on the rank versus adding a constant? I don't feel comfortable with adding a constant, taking into account we need to prespecify things.

The second point -- I'll try to be brief for time's sake. In terms of combining studies with the Bayesian approach you have, in approving a drug, there is replication of evidence or finding. Definitely if we have five studies -- and we could have two studies make it out of five, as you are aware, it would be multiplicity adjustments when we have several studies and you have only two studies, they make it out of the five. By going and

putting those in a Bayesian approach, which we could talk about what's the prior distribution you are using -- I mean, there's a lot of generation. We will be missing that part of replication of a study finding. So I'd like to see the evidence and a study base.

Of course, there are things about how the study size, number of centers. There is much more detail. But I'll stop here if you could address those. Thank you.

DR. BERRY: Thank you. Excellent questions.

The prespecifying the analysis, I agree. The point I made was that you must prespecify the analysis. But that can be certainly a process. So you say this is what I'm going to do, and the data are going to dictate the value of C. The point is that if the company does it and comes up with C equals 29 and the FDA does it, they better come up with C equals 29. This is a process that is dictated by the analysis.

It's comparable to a t-test. If you do a t-test, there's a value of the standard deviation in the t-test and you can't say in advance what the standard deviation is. It's going to be dictated by the data.

It's exactly the same point here. The analysis is specified. As I indicated, the false positive rate is guaranteed. So I can't imagine that there would be any problem with doing that from a statistical or other

perspective.

You say rank transformations and the comparison. Indeed, for reasons that I indicated, I like rank transformations. I gave you a list of things that you can do that are reasonable to do. If you want to see a comparison of the two, the paper that I mentioned in Biometrics does quite an extensive comparison of the rank transformation and the log transformation with C dictated by the data.

In terms of combining studies and missing replication in the two of five, I can't imagine anyone takes this two of five too seriously. Excuse me, FDA. It would be silly to say you have two of five that show a benefit and three of five show it going in the opposite direction so that if you take the totality of the data, it points to no effect. Maybe this is the Bayesian in me speaking -- the scientist in me speaking. You've got to consider all of the data. So it's a mistake to look at just the two most favorable.

In terms of the two of five and needing confirmation, in fact the confirmation is built into the Bayesian analysis. It's something that we will appreciate as we do more of this. By the way, there are people doing more of this, including the Center for Drugs which recently approved something based on a Bayesian analysis exactly

along the lines that I showed here. So I think it's better than the two of five.

DR. STERN: I had two quick comments. I think your presentation was very helpful, and I think the issue of outliers is an interesting one. I think what you presented in a certain sense is another nail in the coffin of the idea of using percent changes. We heard a lot of reasons yesterday where at least there seems to be a lot of distortion in understanding what's happening clinically by using percent changes in this particular metric, that is, acne counts.

But I would say that when you talk about either counts, change in counts, or some transformation in counts, in fact outliers can be among your most interesting patients to look at in clinical medicine. If you have outliers either for the very good or very bad -- and usually studies are clearly much more highly powered than 25 individuals. So if it was a 250 person study, one outlier wouldn't matter statistically within the realm of possibility. But in those, if you have a cluster of outliers, sometimes it tells you about something that for a subgroup of patients is very good or very bad about the therapy.

So although they present analytic problems, in any study they, first of all, point you out -- like we

heard yesterday, is it someone checked off the wrong box kind of error, going back to the integrity of the data. And secondly, in clinical medicine, we often think about, gee, why are these few persons so different from everyone else. What is it about it that was so wonderful or so awful in terms of their behavior under this drug?

So those were two points I wanted to make.

DR. BERRY: I completely agree. Sometimes the outlier, even in the big study, is the important thing. You would completely rule out a medication, for example, if you had 1,000 patients and one of them died of acne. I suppose you could, couldn't you, if you got lots of acne all over your body and it led to death? That would be the nail in the coffin of that drug.

DR. STERN: Yes, Dr. Ten Have.

DR. TEN HAVE: Thank you very much for the very clear and helpful, informative talk, Dr. Berry.

I have a question about the interpretation of transformed data, specifically the log transformed data. You mentioned you could transform back, but you do run into problems when you start looking at means of transformed data and transforming back because of Jensen's inequality and issues like that. And you have the additional problem of looking at differences in logs and then taking means of differences and then transforming back.

How do you handle that when you're looking beyond the p value and trying to interpret a mean effect? And with those issues, have you ever considered a generalized linear model, specifically the log linear model, as a potential alternative where you do have some interpretation issues but at least you know what you're interpreting?

DR. BERRY: Yes. You asked a question but implicit in the question are the answers.

DR. TEN HAVE: One possible answer. I'm not claiming it's the only answer.

DR. BERRY: No. I think it's extremely good.

In terms of transforming back, I do a couple of things. Let's say you're doing percent change. One is to give the percent change, but to do the statistical analysis in a different scale. And it's absolutely standard in the way we do things.

In oncology, for example, where we look at relative risk reduction, it's tremendously important from a statistical perspective, but relative risk has no direct impact on a patient's decision. You want to look at absolute risk, but absolute risk is very difficult to analyze. So you do the analysis in relative risk and give the interpretation in terms of absolute risk.

It's the same sort of thing here. I would do

the transforming back. These are the raw percent reductions, and even though it doesn't look like from the percent reductions that there's a statistical benefit, the transformation gives you a statistical benefit.

I write papers in which I give Kaplan-Meier curves of survival, and the way I calculate the p values is to do a multivariate analysis incorporating all sorts of things that are not in the picture. The p value that goes on for the treatment benefit is in that more sophisticated analysis, but the Kaplan-Meier curves show the survival over time unadulterated, unadjusted for any of those things. I think it's the same effect.

DR. TEN HAVE: Although in that specific case you can get standardized Kaplan-Meier estimates. But there is also the issue of confidence intervals for effects that you'd like to have, but that's an issue we all face.

DR. BERRY: Right.

DR. STERN: Thank you very much, Dr. Berry.

DR. BERRY: Thank you.

DR. STERN: The open public hearing is open for anyone else who would like to come forward. Would anyone else like to speak?

(No response.)

DR. STERN: Then we'll close the open public meeting and go on to the committee -- this is a bit

different than most committee meetings that at least I've been part of in that usually there's a matter of specific questions with yes/no answers, and these rather are specific questions where we're trying to provide guidance and a range of responses and opinions.

So I thought I might try, at least for the first question and see how it works, a little bit in the sense of modification on the usual way which is to go around in a somewhat random order among committee members and have them give their opinions about that question and then have each succeeding person, if they just agree with the prior opinion about it, say yes, they agree and why particularly they think it's more or less important; if they disagree, what their opinion is. And then at the end, I will take the chair's prerogative of adding my own two cents in at the end.

Yes, sir.

DR. KILPATRICK: First of all, Dr. Stern, I apologize for being late this morning. I didn't get the update from 8:30 to 8 o'clock.

Secondly, are we going to review the questions in the order printed or are you going to change that order?

DR. STERN: No. Question 4, which on my sheet was question 3/2, because there was no 4, will now become question 1.

Does anyone else feel we should change the order other than that one?

DR. KING: I think we ought to make question first 3 number 5. I think we can go through by making number 4 number 1, then go 2, 3. So the investigator global severity scale would actually be after you decide about inflammatory and lesion count analysis so you will be able to focus on the co-primary lesion counts and acne types.

DR. STERN: So, 4, 2, 3, 1. Is that the order you're proposing?

DR. KING: I'm proposing 4, 2, 3, 5, 1.

DR. STERN: Okay. Is everyone okay with that? Let's see if I can get it right.

DR. TEN HAVE: Can I make a quick comment or a question here? Bob O'Neill from the -- is it Center for Statistics?

DR. ALOSH: The office director for biostatistics.

DR. TEN HAVE: He came up to us yesterday, a group of statisticians, and asked about actually the first question, should the current success criteria using the co-primary endpoints be retained. He actually had a totally different point of view from what we had yesterday in discussing that question, and actually potentially

combining both the lesion count and the global evaluation into one outcome in the sense that the global evaluation would give some sort of decision and then you would do the lesion count based on a clinical decision made on the evaluation.

He used an analogy from another area of research, and the only analogy I can think of is from psychiatry where you have different measures of depression. You have a clinical evaluation of depression and then refine that evaluation using a quantitative score like the lesion count. So I think that's where he's coming from. You have the physician do a clinical evaluation and then refine that evaluation with a quantitative scale such as the lesion count.

I'm just introducing this because it was introduced last night by Bob O'Neill, but we can leave it for another time to discuss because it sounds like this is an ongoing process. But I think I just wanted to throw that out as where I think some people were coming from when they laid this order out.

DR. STERN: Could I ask a question possibly relevant to that? I've not ever been an investigator for an acne study, but might it be true that in some studies the individual doing the counts might be different than the individual who does the global evaluation, that a study

nurse does the counts and the physician who signs off on it does the global evaluation which would be a bit different since it's two objective measures rather than a test and a single clinician?

DR. WILKIN: They do not necessarily need to be the same person, yes.

DR. STERN: That might make it a little bit different than the example you gave if there's one individual doing the counts and one doing the global evaluation.

DR. TEN HAVE: Right. And I'm not sure what his purpose was, potentially trying to make it more relevant to what happens in practice where you have the same individual possibly doing the diagnosis.

DR. STERN: I'm sorry. Dr. Tan.

DR. TAN: I just want to add very briefly. This is just a way to make the clinical assessment more relevant because we have talked about global assessments and not just the -- it's agreeable with the inflammatory count, but it doesn't catch the non-inflammatory part. So I think the idea is just to combine the information to come up with one endpoint, but we don't know exactly how that should be done. There should be more discussion on that.

DR. STERN: I think some of the things, if there is to be a global assessment, how to make it a

robust, independent, and clinically relevant assessment are probably some of the issues that the committee will come up with suggestions about how to do.

DR. TAN: Right. You can combine with the digital technology combined with the lesion counts, a logarithm, you know, a decision rule.

DR. STERN: So let me start with the question that will be considered first this morning, old question 4. Should acne lesion types (inflammatory or non-inflammatory) be medically acceptable indications?

The questions here got into two areas which I think are important to separate. One, should people put forward as an advanced hypothesis and only be required to meet the need of one, or alternatively, should meeting any of them, even if only meeting one therapeutic endpoint were in fact statistically significant -- would that be grounds for approval for the broad indication of acne?

Or from a clinical perspective, should there be studies that say we have an agent for inflammatory acne, we have an agent for acne vulgaris of mild to moderate degree, we have an agent for comedonal acne and that basically the endpoint or endpoints be determined according to what you're asking for? And if you ask for acne vulgaris and you only make it for comedones, you don't get approval because you've not met the test of what you're agent

required; whereas if you ask for comedonal acne, you make it comedonal, that's all you need to do. Obviously, this discussion is in a sense independent of the discussion we'll have of global endpoints.

Who would like to go first? Dr. Raimer.

DR. RAIMER: I would be in favor, I think, of companies actually going after one indication or the other. I think if they meet either in the inflammatory or the non-inflammatory, comedonal acne, if they actually show improvement in any one of those categories, they should be approved for that specific indication.

I guess they should only be approved for acne vulgaris in totality if they actually improve in both categories. Otherwise, it should be for one or the other, and I think it should be acceptable to be approved for inflammatory or for non-inflammatory acne.

DR. STERN: Perhaps we should ask one of our statisticians what if they go for both and only make one. Should that be non-approvable? Should we think about, in fact, doing a Bonferroni or some other adjustment for multiple comparisons? Would you change the standard, or would you just say you asked for X and you didn't meet the bar for that, so you've got to start over if you want to it for comedonal acne or inflammatory acne only? What would be your suggestions?

DR. TAN: If the protocol is specifically for inflammatory, that would be just for inflammatory. Of course, I think they probably do for the secondary, maybe in the protocol if they include data for the non-inflammatory, but they cannot because the criteria will be different. So I think it's a non-issue. It will be just inflammatory.

DR. STERN: But my question is someone decides to go for what now is the usual, if not the only, indication that we have in this class of agents, which is acne vulgaris mild to moderate, and there's an analysis of response in terms of comedones statistically significant and in total because, as we saw yesterday, it's possible for one of the two arms to drive it, but they don't make it for inflammatory. The example we saw yesterday where the difference was 3 with the standard deviation of 2. If the difference had been 1 less, if it had been 2, it probably would not have made significance. So then in that particular instance, total would have probably still made significance. Comedonal would have been significance, but inflammatory wouldn't.

They've said, we have a product for acne. Do you say, well, we're going to have to test you for significance according to the criteria of you've done three comparisons, so the p value now has to be basically the

equivalent of .025 rather than .05 by a Bonferroni or whatever people want to use, or do you just say, I'm sorry, you put this product forward to really treat acne in its totality and you're only treating one element of it. You have a non-approvable product.

DR. TAN: In this case they should adjust it for Bonferroni. Multiple comparisons.

DR. STERN: And then what would you do in terms of labeling?

DR. TAN: I think you just have to report actually what happened.

DR. STERN: So you would say, okay -- we haven't even recommended about whether we want to continue to even have the total count, whether that's a useful addition. But under this current system with that agent, the agent we saw yesterday that made a 1 fewer inflammatory lesion difference, so the p was greater than .05 for that test, but it looked like it probably would have made even a p of less than .025 for each of the other two because of the substantial difference in the change in comedonal lesions, you would say even though you went in for acne vulgaris, we'll give you comedonal acne only because you've made it under the more stringent p criteria. It's a significant effect for that.

DR. TAN: There's confusion here. When they

conducted the trial, all this should be specified in advance, the decision rule.

DR. STERN: Right. I think you're right. But let's assume that they said that we're going to show you that we'll make both better, and they only make one better but they make it a lot better, one component of acne.

DR. TAN: Then that's a tough decision. There are examples at CDER where they have run into this situation. That's a very large discussion to reach such a conclusion like that.

DR. TEN HAVE: Can I ask a question of the FDA? Are there analogous situations in other contexts where you do run into this situation where you have a primary endpoint that maybe may not show significance, but there are secondary endpoints that do? In a way this is probably a similar situation. How does the FDA handle that?

DR. BULL: Clearly this comes up. One of the typical ways that this is addressed is that it's viewed as more hypothesis generating, and the sponsor is asked to go back and do further work. But my experience has been that this is not typical that an indication would be carved out unless the evidence was really compelling. I mean, it would really have to be very persuasive data that, in essence, would say that probably the wrong primary endpoint had been identified and that all of the data, the totality

of the data, within the study really was supportive of using an endpoint that was not the prespecified primary endpoint.

But the usual and, I would say, the most common way of addressing this would be that you have something that -- studies always give you more information than you initially thought you were asking for, and as we enter into that uncharted territory of unanticipated consequences, when you look critically at different elements that you have under investigation, that the things that you find out can lead to an evolution of thought to other ways that may necessitate additional clinical work. But the usual recommendation is additional clinical studies.

DR. STERN: Just one other aspect of the question that just was asked. Are there studies with co-primary endpoints? For example, a drug that would be for both depression and sleep and, let's say, when you test it, had no effect on depression compared to placebo but it was wonderful in terms of helping people sleep and wake up rested.

DR. BULL: I think the original question was about a secondary endpoint, but here if you've got two co-primaries, I would say there would be the possibility if it was truly a prespecified co-primary -- they weren't interdependent co-primaries -- that it would be a

possibility of that being a carved-out indication.

DR. TEN HAVE: So what would this be called? This situation. Would this be two co-primaries or one primary and one secondary?

DR. KING: I guess I have the fundamental problem about co-primaries and primaries. It seems to me if you make an hypothesis with two parts and one hypothesis fails, it just fails.

The reverse question is if you have two elements, such as depression and sleep or papular or comedonal acne or whatever, it seems to me that if you went in for a specific hypothesis and didn't make it, you fail. If you went in for the total one and you had two parts, it seems to me that you'd either have to have co-primaries or the sponsor would have to come back with an amended proposal and it would be reevaluated on that basis. So you either have co-primaries or you failed one of two and you have to submit an amended request. Is that not what's happening now?

The question is, if you went in for acne vulgaris and you made it on comedonal but not on inflammatory or vice versa, does the sponsor have to submit a revised application or is that analyzed in toto?

DR. WILKIN: Well, how the committee is approaching this now is -- I think I'm hearing that maybe

you're not going to recommend looking at total, that you're looking at inflammatory lesions and non-inflammatory lesions separately.

I can tell you what has happened. If you are going to recommend that to get the indication acne vulgaris without limitation to a specific lesion type and you're going to say that to earn that indication, you have to demonstrate significant changes in both inflammatory and non-inflammatory lesions, you will be asking for a higher efficacy standard than we have asked for in the past for the indication acne vulgaris because what the history has been is literally one lesion type has been allowed to drive a win in that lesion type and total.

Dr. Fraser had a very articulate presentation yesterday morning about some of the difficulties when you actually have a positive in a particular lesion type but because the comedonal lesions are more numerous and there's more variability, sometimes that gets lost and you don't get the good p value for the total.

But the products that are out there are the products for acne vulgaris have won on total and one or the other of the lesion types. If you're going to ask for both inflammatory and non-inflammatory, again that would be a higher standard than what we have asked for.

DR. STERN: I think the issue should be what

are clinical relevant standards. It seems to me that we saw yesterday that a large proportion of acne is, in fact, treated by specialists, and we would hope that people who are not specialists in dermatology who treat a fair amount of acne pay attention to how it's treated. To my mind the approach to inflammatory and comedonal acne is often different. I thought I had some interest in following the literature on acne therapy, and I realize sometimes I'm not sure which agents are in fact only helpful in comedonal or helpful in inflammatory based on both the published literature, which really follows the clinical studies, and the labeling.

So to me it seems that there are two pretty distinct endpoints that any reasonable clinician and in fact almost all patients understand as different manifestations. It's great to have one product fits all and that's a product for acne vulgaris, but it's also at times helpful to have one product that mainly is targeted against one or the other. At least to me it seems sort of logical.

And then the following logic is if you go for the bigger jackpot and you would have had a hand that would have won the smaller jackpot, I don't think you still get the smaller jackpot in Las Vegas or other places. I don't know the terminology because I don't gamble, but that's

sort of the analogy.

So I think you have to play by the rules that you set, as you said, Lloyd, when you go forward with your product. It's not as if these were coming out of the chemist's hands and hadn't had phase II trials, hadn't had basic science, animal models, phase II trials that would give you, the sponsor, some good idea about where this should be working. I mean, we don't go to phase III trials immediately. But at some point you have to decide whether you're going to go after the big jackpot or go after the smaller jackpot and just have one indication.

DR. KATZ: It seems like we're all in agreement, including industry. In view of Dr. Fraser's presentation yesterday, it sounds as if they would like to separate them also and it would make it very much more objective, objective for the clinical studies and how the practitioner would be able to evaluate and treat patients. It sounds like industry would be in favor of that too.

DR. STERN: Dr. Plott?

DR. PLOTT: Yes, I think that industry would welcome an opportunity to separate inflammatory and non-inflammatory lesions. I think for lesion counts, as we're addressing it, inflammatory lesions are more of a concern because they make up a minority of the total lesion count. And for inflammatory lesions to also get total lesion

counts can be more difficult because of that disproportion.

DR. STERN: Please correct me if I'm wrong. I think I've heard from the committee that we think the total is really irrelevant, that it is two separate tests, that you test for change in inflammatory lesions, you test for change in comedones/non-inflammatory lesions. If you bring forward a product that you say should be approved for acne vulgaris, if it passes each test, that is, both tests, then it should be approvable, assuming its safety, et cetera are all reasonable. That should be proof of efficacy for acne vulgaris. If you come forward with a hypothesis that this is something for inflammatory acne, you're only subject to looking at inflammatory lesion counts and correspondingly for comedonal, and that basically taking the total counts and adding them all together is really irrelevant in terms of clinical decision making. So you don't have to meet that bar because it's not going to be there anymore, if I understand what the committee thinks would be the most reasonable way of looking at these kind of studies. Please correct me.

DR. KING: I guess it gets back to first principles. What is acne vulgaris? I guess there's a debate on what is the meaning of "is."

(Laughter.)

DR. KING: I think if you say acne vulgaris and

you only have the phase where it's primary comedonal, it is acne vulgaris. If you have acne vulgaris with primary inflammatory lesions and not much comedonal, which can happen too, that is acne vulgaris. So I have difficulty dealing with you have to have both simultaneously where one could be numerous and not so apparent, the other can have a few red ones and it's a big deal.

So I'm in favor of separating them, and I think the marketplace and prescribing habits will determine whether something that's primarily inflammatory or non-inflammatory has adjunctive or actually is a total treatment. It seems to me that we ought to judge them on their merits.

Conversely, if you go in for one and you flunk, you flunk. I don't see how you can get around that issue.

DR. PLOTT: Let me just restate what I heard. You've taken the total lesion count and set that aside. You've said that it's not necessary to win there. But to get an acne vulgaris indication, you need to have inflammatory and non-inflammatory. You need to win in those lesion counts.

DR. STERN: I actually, as usual, misspoke. I think, as I understand Dr. King, all of this is acne vulgaris and it would be acne vulgaris comedonal and inflammatory, if you went in for both and won both. If you

went in for inflammatory acne vulgaris and you made that test on the basis of inflammatory lesions, your indication would just be inflammatory acne vulgaris. If you went in for comedonal or non-inflammatory acne vulgaris alone, that would be your indication. But in order to have it say in your labeling -- I don't know whether they go before or after, but inflammatory and comedonal acne vulgaris, you would have to pass both tests.

Correspondingly, if you only passed one test, when you asked for that indication, that would mean that your product didn't make it on the basis of those tests because you had gone for a home run and you had only gotten a two-base hit. You had said I'm going for a home run and you didn't get there.

DR. PLOTT: I'd agree.

DR. STERN: I think we have consensus on that, I hope.

DR. BULL: Excuse me. I would just like a clarification.

DR. STERN: Sure.

DR. BULL: Are we hearing that, let's say if the home run is that you get both the inflammatory and the non-inflammatory, if you get to second base on either/or, you're saying that's okay? Is that logical for clinical practice?

DR. STERN: I'm saying -- and I bet the statisticians here would agree with me -- that advance hypotheses are the key here and that a sponsor should be required to say at the time when they sit down with you to design the phase III study, this is a product for acne vulgaris either inflammatory only, comedonal only, or both. And if they go for both, they're going for the home run. They're going for meeting both tests, significant improvement in inflammatory lesions and independent significant improvement in comedonal lesions.

If they only go for one, they may or may not want to also do the other part of it, and then they can, obviously, publish papers and do things, just as there are a lot of things that don't go in the labeling that come out of clinical studies that get published in the literature that allow the practitioner to understand how drugs are used outside the labeling. But you can't then go back at the end and say, oh, look, we also made it for comedonal, so label us for both without doing separate tests. You can't go for the low and game it to high, but you certainly can then publish a paper that shows this stuff, not only did it work well for inflammatory acne, but look at this study that showed a significant difference in improvement for comedonal acne, but it wouldn't be part of your labeling without additional studies.

And I hope I've represented people fairly in that.

DR. TAN: I think so, yes.

DR. STERN: My sternest critic, Dr. Kilpatrick.

DR. KILPATRICK: Dr. Stern, as you can tell, I'm in a rather different mood this morning.

(Laughter.)

DR. KILPATRICK: But I just want to reiterate what has been said. You have to hold the feet to the fire, depending on what was decided beforehand, and that has to be made very explicit. The political issue is for the division to cater to the sponsor when they don't meet the criteria.

DR. STERN: So let's go on to question 2, which is still question 2. How should lesion counts be analyzed?

Would one of the biostatisticians volunteer to take a first --

DR. KING: I thought we were going co-primary for number 2. We were just evolving from acne versus that, then co-primaries. The same question.

DR. STERN: Lloyd, I'm sorry.

DR. KING: I think I'll just start off with that. Should current success criteria using co-primary endpoints be retained? I think that it really follows from what we just said that if you're going to go for both

indications, you're going to have an element here of one, two, or three things. So in that sense, it's an evolutionary of that concept, so I think it should be retained for that part of the study application, just prespecified. So it should be retained for that.

DR. STERN: I think the other part of co-primary, if I understood correctly, should we keep the global? Does the global add anything? Here I think we heard some very interesting things yesterday.

We heard from Dr. Alosh when you use the counts, you could, in fact, have models that at least as the co-primary endpoint of success, as it's been used in at least the clinical trial he presented, really made it irrelevant because it was basically driven by the counts or change in counts in those two models.

However, I think we also heard some things from a variety of the acneologists yesterday that now technology has moved forward that for those of us who can't even remember the order of questions, let alone how a patient looked 12 weeks ago, that there are aids to memory either for the individual or ways of gathering data that allow, in fact, a more independent, clinically relevant assessment of an individual.

So I guess I would say as it's been used so far, at least in this one trial that Dr. Alosh talked

about, the current co-primary, as it's been done, seemed to be so driven by the results that were there quantitatively as to add little.

But to me we heard a tremendous amount yesterday about new methodologies that would really allow an independent observer either working for a company or independently, as Dr. Plott implied yesterday, to robustly, in the most important way, particularly if they're presented out of order -- and they're all photographs -- really judge, hey, does this person really look substantially better with a training scale, et cetera.

So I think the issue is should there be something about how often people are really a substantial amount better on clinical criteria. That to me as a physician is a no-brainer. But what it should be and how it should be engineered is much more difficult.

DR. KILPATRICK: This morning, Dr. Stern, I'm going to agree with you and repeat, in effect, what you've said because I started yesterday by, as a statistician, being very attracted to the concept of counts because of the numerosity, being attracted to the technical developments, namely photographs, where we can, as I said, possibly get other measures from the photographs such as severity in terms of color, density, et cetera. But as I said yesterday again, the problem then is how to combine

those things and make it meaningful.

And in the process of listening to the experts yesterday, I've come to the conclusion that we have that facility in the hands and in the minds of the experienced dermatologists or other physicians who look at photographs. My point is that I think that I'm taking Wilma Bergfeld's stricture that we should have a simple, one outcome whether it's for inflammatory or non-inflammatory and try to heed her recommendation.

So I am going away now from counts towards the IGE or some version of the IGE. I think it should be more than just success/failure. I think it should be more, probably a 5- or 6-point scale. What that is I don't know, but it should incorporate some of the scales we have seen and the information from the photographs in terms of counts. Whether those be counted by impression or by literally counting, that's another issue. But I feel very strongly. And all I'm doing is repeating what you're saying, I think, by saying we should go to a modified IGE which tries to bring all of this together.

DR. KATZ: Basically when we're talking about global anyway, when somebody is judging whether it's from the patient or from the picture, which the picture would be more accurate perhaps, we're counting. They're subliminally counting. They're not deciding that a patient

is better or not better by just some spiritual feeling. I mean, basically even if you say you're not counting, you see somebody with five nodules, and then you see somebody with one, you're saying, I don't need to count basically. This patient is better. But it's better because you saw five before and one now. And you just have that impression even if you're not counting because it's there.

So that's why the global is not -- I don't even know what that means because when you say global, the patient is better no matter how the technology. The technology just makes it more accurate.

Although, granted, it's a possibility what Sharon said yesterday, that counting is not perfect -- and it isn't because you wouldn't count 20 lesions that were 5 millimeters compared to 20 that were 2 millimeters and the count would be the same. That doesn't really happen in real life. I mean, patients with acne don't get better because they have 20 and the 20 look better. That may be questionable. But the truth has its day because if they're getting better, that 20 becomes 5 or 2 or 10. So they don't just keep getting better by volume. That doesn't happen.

It does with comedones. So if they have a forehead full of comedones and you count 200 and they're better, basically you're counting 100. You do it more

accurately with a picture.

So the global, both from the statistician and from Dr. Fraser's presentation from Stiefel -- I don't know that global has any additional factor and may be obfuscating.

DR. STERN: Dr. King.

DR. KING: As we've had the question number 2 and so forth, actually the discussion here has touched on question 3, how should they be analyzed. So my response to revised question 2, should the co-primary endpoints be retained, the answer is, in my mind, still yes with the provision there's going to be an evolutionary process where now with the new technology you can validate what you see. Visual prejudice regardless is still prejudice.

So I think that the agency is still left with how do you validate the data. An experienced numerologist or acneologist or whatever -- never mind that in the sense of rules of law and the FDA. I think we have to retain this for now and have some studies showing that you can validate what you're saying you see. Human error, after all, is still human error and I don't care what prognosticators or acneologists say. There are problems with this that I think the agency has to deal with. So I'd like to see us continue the co-primaries and see if we can improve the technology to validate what we all think we're

seeing.

DR. SAWADA: Dr. Stern, I would agree with Dr. King. I would say retain the co-primary endpoints. From my view, the thought is if you do the investigator specific evaluation or the global, the setting up of the criteria, a 5- or 6-point scale, leave it to you guys to figure out what your scale is going to incorporate. But photographs is great. The nice thing about it is you could have two separate evaluators, one counting lesions and the other one just doing an overall assessment into which category this falls. I think that just gives you a little bit more information when you go in and do your statistical analysis and also leaves for a blinding category or something of that nature that you wouldn't get the two areas confused.

DR. STERN: Other comments? I'm sorry.

DR. KILPATRICK: There's one other aspect which we haven't touched on and which I think we, in effect, agreed on yesterday, and that is the variability of the disease under treatment. I thought it was agreed yesterday that it would be useful to have possibly not only baseline but at least two examination points, perhaps at 6 and 12 weeks. That could be effected under these schemes by having evaluations from photographs or from doctor visits of the progress of the disease at 6 and 8 and 12 weeks, but the implication for my mind is that the clinician would be

asked to come to a decision on one 5-point scale at the end of the 12 weeks, let's say, as to what category the patient fell in, not two separate ones. That's again an attempt to make things simpler.

DR. PLOTT: From industry, I think there's not such a concern about co-primaries. I think that most firms feel that if a product would be worthwhile, it should make a difference and have a clinically meaningful difference

The question is more what is the global. What's the right global to use? And maybe that's another question. When we get to that, I'll address that more. But I think from industry it's more of a concern about what the correct global is.

DR. STERN: So I think we've pretty much come to a conclusion that the -- I'm sorry.

DR. WILKIN: I didn't mean to interrupt. Actually you may be giving the answer to my question. Shall I ask? Okay.

I think we're hearing that there's utility in retaining the global, but in the previous question, you've opened it up now, which is consistent incidentally with our Code of Federal Regulations, 201.57(c) which talks about the indication section of labeling. The indication doesn't have to be for a specific disease. It can be for specific aspects of a disease. I think that's quite consistent with

having indications that might be inflammatory lesions of acne vulgaris, non-inflammatory lesions of acne vulgaris, and then another which would say acne vulgaris both inflammatory and non-inflammatory lesions. So far, I think that really works well.

Then you went to a question, which was originally question 1, and you're talking about global. And my question back to you is are we now hearing about global for that third indication which is acne vulgaris both inflammatory and non-inflammatory lesions, or are we hearing that global will also play a role in those other two indications? Will global play a role in inflammatory lesions and also the non-inflammatory lesions indication?

Here is one of the things -- if I could just remind the committee, our experts, the acne numerologists, as Dr. King refers to them, I think they were telling us that the inflammatory lesions ultimately drive a lot of the message that comes from the global. And Dr. Alosh did some looking into two NDAs that we have and I think basically it's about 4 to 1, the effect of the inflammatory lesion compared to the non-inflammatory lesion on global.

So if we could get some clarification on where global fits. Is it only for the acne vulgaris, both inflammatory and non-inflammatory lesions, or are you also recommending global play some sort of role in the other?

DR. STERN: I think you've shown that decision making is never clear-cut because under the way we've done it, unless you have global as -- I guess there are two questions about global. It seems to me for inflammatory and both indications, the global by photographs is reasonable to require because the inflammatory lesions -- again, we're talking about current technology, current photographic technology. So if you don't go to something where you're extracting comedones or some other kind of very sophisticated, not usual measure, that's pretty straightforward. It seems to me for inflammatory and for the dual indication, global has to be part of it.

To me, because of the difficulty of seeing comedones on photography, you wonder if that should be a criteria the other way, that there's not significant worsening because let's look at the product if we say what do we do about the product that does a wonderful job with comedones, gets rid of 100 percent of them, but doubles the amount of inflammatory acne in these patients? Now, to my mind, that's not a product that should be approved because probably those individuals on their photographs would be likely to look worse, not better.

So I wonder if this is kind of a measurement problem and whether for comedonal acne you, at least at the current time, until there's a better technology that people

agree really capture that, you still require the same photographic standards but the standard is being not significantly worse than baseline in the global evaluation.

DR. KILPATRICK: I'm showing my ignorance again, sir. In the situation you describe, would the aggravation of inflammatory lesions not be an adverse side effect?

DR. STERN: I'd like to ask Dr. Wilkin or one of his colleagues.

DR. WILKIN: Well, we actually thought we had that taken care of in the past by asking for total lesions. I mean, that was part of what the total lesions was, and it really was the other way around. It was, if you had a product that worked well on inflammatory lesions, to make sure there was not a comedogenic ingredient in the vehicle.

We're hearing I think the no loss on global. Whether a lesion is inflamed or whether it's not really particularly inflamed, an inflammatory lesion, the lesion count is still going to be the same. Maybe that is a way to discriminate between something that is a side effect and something that is actually a benefit.

Now, remember that many of the local intolerance reactions for these products are going to be seen somewhat earlier on. This evaluation you see is going to be out at 8 weeks for those who are still standing in

the trial, the people that are not having such severe local intolerance reactions that they want to continue on. So by that time, it's a meaningful question. Actually it would be nice to hear how we should think of that.

DR. STERN: Dr. Katz?

DR. KATZ: I may not get it. I don't see what global adds to anything. It seems to me like using the term confuses it because it's very subjective. So when people are evaluating the drug, it confuses the matter as much as the total count would. I mean, when we look at somebody and say they in general I think this person is better, you're doing it on the basis, either by photography or in person, of counting lesions once again. Am I not right?

DR. STERN: My response to that would be if we had a 5- or 6-point scale that was photographically well defined, perhaps even of people of a different gender and different skin color, and you had clinical data that told you, first of all, two things -- one, we did a trial of -- you know, mild to moderate in the 6-point scale covered basically 4 points of that 6-point scale before you got to severe -- that 80 percent of the people were level 2 acne and 20 percent were level 3, and on average they went down by 1, that would tell me something about really who the product had been used in, how much response, and I'd view

it very differently than another clinical trial where 50 percent were 3's and 4's and they went down by 2.

So when I start to think about counts -- when you get to very large numbers of counts, you know that's a heck of a lot acne. But when I try to picture someone says it went from 25 lesions to 15 lesions -- you know, we were seeing yesterday drugs that were approved with an average change from 18 to 15 inflammatory lesions or maybe it was 21 to 18. I mean, that level of improvement.

I sort of asked myself, so what would be different in these pictures? And I know it was statistically significant, but what I really want to know is what are the odds that a patient really would go down from being at one clinical level to really at least enough better that you could tell with two photographs.

So that's why I like the global if it's a standardized global. It both tells me about the treated population. To say about the treated population their mean number of inflammatory lesions was 27 at baseline, plus or minus 11, that doesn't tell me as much as half of them were 2's and half of them were 3's and there were no 1's and 4's in this 6-point category.

DR. KATZ: But that's even more subjective saying it's from a 3 to a 2. What defines a 3, what defines a 2 you're going to define ahead of time by

numbers, otherwise everybody has their own opinions. I don't know how you would gauge that. I have not been involved in any acne studies myself. Maybe it would be clearer if I had been involved in that.

DR. KING: Let's go back to Dr. Wilkin's original question which is how should we use globals and are we talking about globals only for the complete indication, or are we talking about for the indication for comedonal or inflammatory.

I think Dr. Bergfeld's plea for simplicity comes back to the same thing. If we agree that counting is going to be part of all this, regardless of how you do this, then one of the purposes, at least regulatory, is if you talk about proving efficacy, you're going to have to have some specific parameters, and I think everybody agrees on numbers. But it makes common sense if you're going to talk to people who are going to do these studies that you have to have some global to verify that the number of bumps that went down, that there is a correspondence between numbers that you see and then the global impression. After all, the patient is going to look in the mirror and they're going to do the same thing. They're going to count or not count.

So it seems to me that you have to retain the global for comedonal, a global for inflammatory, and a

global for total. And it seems to me we're trying to separate these are parse it out so that not only can the agency do this, but the statisticians can talk about apples and apples and oranges and oranges. And when you present a study, you're going to define the number of inflammatory lesions, count them, have a global for that. You're going to count the number of comedonals, if you're applying for that, and do that global. And if you're going to go for the whole ball of wax, you're going to do both.

So it seems to me you're parsing or teasing out how the statistician can approach this as there is concordance between the numbers you count and the investigator saying I think this is better or worse, and having at least two people look at photographs will definitely tease out what may be a backup. Having more than one blinded observer is a great thing to do.

DR. KILPATRICK: I want to make two points.

One is I think it's obvious that the word "global" should disappear when it is inappropriate.

The other point. I want to speak against the use of counts as a primary endpoint because I think we had some discussion yesterday here and then about the difference between clinical significance and statistical significance. And it's been demonstrated adequately that we can get, as you yourself said, sir, statistically

significant results which have little or no meaning. We're now talking then about the use of counts in some sense to augment this clinical evaluation. Again, I'm in favor of that.

But I wanted to speak to Dr. Katz about the subjectivity of such a multi-level scale. I'm not too concerned about that because I think that with the photographs that we saw or the possibility of showing photographs which are typical of different types of patients, that will facilitate people putting a given patient into a given class. It makes more sense in terms of the label, in terms of the conclusions rather than this other plus or minus 7 comedones or inflammations.

DR. PLOTT: I wanted to also speak to how the studies are conducted, and I believe that often most firms instruct the evaluator to go in and do the global evaluation first so that they are not biased by a count, if the same investigator does the global. Just how is this patient doing, where do they rate on this scale. As it's been pointed out, almost every study has a different scale, so you have to look and see what that definition is and try to fit that patient to that scale.

Subsequent to that evaluation, then the evaluator takes the time to count each lesion. More and more, I think we're using a system to try to count in

specific areas so that you're not trying to count maybe 200 spots across an entire face. The system that Anne Lucky has published is probably the most commonly used one across most of these studies where inflammatory and non-inflammatory lesions are counted and sometimes even subdivided into papules, pustules, open and closed comedones, which becomes enormously tedious.

But that's just for information how that's done and why a global could be useful in evaluating just what's your doorway impression of the status, rather than improvement, of the patient at your first glance.

DR. KATZ: How is the particular scale defined? When the investigator is told, well, just give us a 2 or a 3 scale, 4, how is that defined, 2, 3, 4, whatever?

DR. PLOTT: And that's really the reason that we're here today, to figure out what's the best scale to use. Every scale is a little different, and it's defined by the companies.

DR. KATZ: No. I understand, but what do they say? Put it in the 2 scale or 4 scale dependent on what? You ask them just tell us how many non-inflammatory lesions are there, or how do you define that?

DR. PLOTT: The scales that Dr. Carr presented yesterday are some examples. A 1 is usually defined and the better that definition is, the better that scale could

be --

DR. KATZ: No. But what will you tell the person in some -- what I'm getting at is, do you tell them that more than 2 lesions or more than 5 lesions are inflammatory, non-inflammatory?

DR. PLOTT: Well, in my opinion it's best not to tie it to a lesion count rather than to say, okay, inflammatory lesions persist or predominate, non-inflammatory lesions predominate.

DR. KATZ: But basically they're counting. How would they remember 4 weeks later? How is it possible?

DR. PLOTT: Hopefully it's a status score.

DR. STERN: Dr. Katz, I think one of the things that the committee has an opportunity to advise or comment on is the point you made that the current flexibility or perhaps even capricious nature of the scales and the lack of standardization and the lack of memory jogging makes them, as you point out, not very useful. I think one of the things that we can recommend is what are the things that make it useful for both standardization to make it an even playing field and also to really help people jog their memories and help for independent evaluations. We heard some suggestions about scales that are basically photographically based. There are ways of developing such scales, and rather than, as you implied, each company kind

of coming up, well, this time the way you decide on better is A, B, and C, and it's a 4-point scale, rather saying these are the standards by which you will categorize patients and these are what you will have when you view a patient 12 weeks after entry in terms of their initial appearance, that you'll be able to decide whether they got better.

So I agree with you fully. The way the global measure is used now, there's not a measure. It's kind of the measure of the day or the measure of the company without any way of reproducing it.

I guess some of what I think I heard is that some people believe technology has moved where we can both help the investigator and in a sense have the most important record for dermatologic evaluations which is for an independent evaluator to be able to go and say, oh, this is how your patients looked at start and this is how they looked at the end, or better still, have them presented randomly. This photograph is better than this one by this degree and you don't even know which is before and after.

That to me is the ultimate test for products like this that are meant to improve the appearance of an individual, and in my opinion it would be a shame not to strongly advise the agency that, gee, it would be nice that if you made sure that you helped industry develop these

technologies and applied them uniformly across studies.

Dr. Tan. I'm sorry.

DR. TAN: Yes. I just want to reiterate several points that have already been made.

I think we have seen enough evidence that the lesion counts are not totally satisfactory because of the weak correlation with the clinical endpoint. In an ideal world, you want to define a clinical endpoint. But here that is the global assessment. The current way is not satisfactory. That's what we all agreed on. But the effort toward to find a good clinical endpoint -- that's what we should strive for in this kind of trial. We want to find a gold standard clinical endpoint.

But if you define that global assessment, you could use the technology where you have a decision rule. That would account for the lesion counts. And if the evidence in the inflammatory lesion counts, for example, is so overwhelming, so you make your decision rule such that it's driven completely by the lesion count in that case. You can do all sorts of things, but the goal is to define, to get a better clinical endpoint, better than the total lesion count that we currently have.

DR. STERN: Lloyd.

DR. KING: No more preaching. I just think that the Academy of Dermatology, since the dermatologists

are the ones who treat most of this, should evolve the pictorial scales and get a buy-in here. We can debate all we want and the FDA can propose all the scales they want, but I think you need an acceptable standard for the consumers and the people who treat them, the dermatologists. It should be something that we should lobby for because I think that the FDA can't make the dermatologists agree. In the story with Accutane, you can put out what you think is safe and efficacious and the right thing for patient rights, but in the end prescriptions are being written by dermatologists and argued about that. So I don't want there to be a feeding frenzy or a big deal saying we did X, Y, and Z in face of opposition by the consumers and the dermatologists.

I think that industry has a right to expect uniformity. They have a right to expect to be treated fairly, and they have a right to present their data. And it's up to the FDA to interpret the data, and without some benchmark of photographs and whatever, I think we're just going to go around in circles.

DR. STERN: Since we've been concentrating on the global scale, with the committee's permission, I thought we might go on to question 3 at this point. I think we've heard a bit about what investigator's global severity scale should be used and I think we've heard

general principles of we're not sure which one. We know that there's not a single one out there that's universally adopted and accepted and it's a goal to get one that meets all the usual tests.

But the second question I don't think we've yet addressed, which is at what level should it be dichotomized into success or non-success. I guess I would say that would include is success other than clear or nearly clear -- does also a change within the scale count as success in therapy, and then obviously how much change depends on the nature of the scale. But I'd leave that open to Dr. Kilpatrick.

DR. KILPATRICK: No, sir. I'm again agreeing with you that I don't think it's necessary, in my terms at least, to dichotomize this 5-point/6-point scale into two levels. It sounds like if we're going to a 5-point scale, the FDA could use logistic regression, continuation ratio, or something like that where you look at the odds from one level to another, and that could be interpreted I think in a meaningful way.

DR. STERN: Other comments on that issue which I think is really driven by however the scale is developed. Dr. Tan.

DR. TAN: Yes, I agree. You need to improve the scale first.

DR. STERN: I agree.

DR. PLOTT: I would agree. The dichotomization to clear or almost clear is a very difficult level to achieve and that there probably may not be a need for dichotomization.

DR. STERN: I think we only need that now because the current scales are so uninformative and unstandardized. But if there's a good scale, then I think we've all said that improvement is what we're looking for and there are scale-dependent ways of testing for that.

DR. PLOTT: Let me speak to the scale. I think that with the current scale, it could be biased, as we've heard, toward inflammatory lesions. Photographic analysis can also be a bias toward inflammatory lesion counts because you don't pick up the subtle comedones. So for that reason, it might be conceivable to think about a scale that is more specific to the type of lesion count, as was suggested yesterday.

DR. STERN: I think Dr. King, if I understood him correctly, suggested that moments ago, and I think that has a lot of face validity to it.

DR. TEN HAVE: Although some of the new digital photography can make contrasts to highlight those subtle comedones. So I think that development of these new photographic techniques is probably improving the

capability of the global assessment to pick up the non-inflammatory comedones. Technology is improving it so that maybe your concerns will be resolved it sounds like, or some of them anyway, in terms of the --

DR. PLOTT: Some of these lesions, dermatologists will tell you, they can really best count if they can feel them. When you speak to these investigators in the meetings, well, you know, these photographs are difficult to count comedones because they almost count them that way.

DR. STERN: Again, at least some of us would be most concerned about products that make a difference in the appearance of the individual. The number of people who are going to say because some barely perceptible bumps on fine palpation have been reduced in number they feel better about their appearance is probably small. So I think one of the things we're always saying is let's ground this in things that really make a difference to the patients who are the people we treat and measures that, as best as we can, reflect what most people would agree is important.

DR. KILPATRICK: Well, sir, you've opened the door to my request, plea that we involve the subject in the evaluation in some sense whether it's in the trial or in the label. Again, I think that we're going towards something. I think again to be perhaps rather critical, I

think that some of the discussion was rather derogatory towards the involvement of the patient. The patients have a very great interest in his or her own appearance and I think can be educated, to some extent, perhaps not with the same accuracy, to follow on through looking at photographs of himself or herself. But that's my one shot, sir.

DR. STERN: Having been a bit involved in some kinds of clinical studies -- and perhaps you're different -- I find that individuals have the hardest time looking at themselves. I've found that people, non-medical professionals, judging improvement or state of disease in other individuals in fact can do a very good job of it. So I think whereas it's been my experience when you ask patients about themselves and how they're doing, it's often a lot of factors other than the objective finding. So I think a non-physician, non-medical personnel's perception of what's better is extraordinarily important. But I find that, for example, with me I don't like to monitor my weight even though that's an objective thing, and how I look in the morning I definitely don't like to monitor or can I say whether I look good or bad because it's more dependent on other factors.

DR. KILPATRICK: On a personal note, I had photographs taken recently for a church directory and I liked none of them.

(Laughter.)

DR. TEN HAVE: Can I just add one thing regarding patient input? One area where patient input has been probably more valuable, not so much in determining outcomes or assessing outcomes, but prioritizing adverse events -- this may be an issue here -- you may get an idea from patients that certain adverse events don't matter that much, just slight irritation or whatever, whereas other adverse events may have more of an impact. I'm very naive here, so I don't know what the issues are. But in psychiatry there's some effort to get more patient input in terms of prioritizing adverse events.

DR. STERN: I think that's extremely important.

MS. KNUDSON: Dr. Stern, I would like to echo very much the patient input into all of this. I'm really appalled at what I've heard about the lack of specificity in so many areas having to do with acne treatment, that the global scales are not sufficiently specific, that the counts have not been necessarily specific, that quality of life hasn't been assessed. All of this is really enormously important in a disease that is so rampant that so much money is spent on treatment of, that we really have to come to much greater specificity before a drug is approved for marketing.

DR. SAWADA: Dr. Stern, I just want to make my

comment too. I think that a very simple question, asking patients in a study is how does your skin feel, is a way in which to assess some of the adverse events. Is it oilier or is it dryer? These are things they can interject. It's more of an essay question than it is a yes/no question. And these things probably could be collated and judged for the companies' benefit as well. So those I think are important inputs in a study.

DR. STERN: Other comments?

(No response.)

DR. STERN: If it's okay with the committee, it seems like we're on a roll. We've taken care of most of the harder questions. Rather than taking a break now, we'll continue on if that's acceptable to the committee. Is that okay?

The question is, how should lesion counts be analyzed? I think we won't hear a lot of votes for percent change as being a well-behaved metric for this particular disease, and I would like to turn to the statisticians about their feeling. Do you just count them, or do you think about transformations, et cetera?

DR. TAN: I think as Dr. Alosh presented yesterday, the percent count reflects one aspect of the clinical efficacy. It's still useful, but because of the variability it introduced, we should be careful. Like the

rank test could be used or log transformation could be used. But whatever test, I don't think you need to have, you know, just one test to be used, but you need to be specific in the protocol exactly what you're going to use. The procedure to analyze the data should be specified.

DR. STERN: You would allow the sponsor to negotiate with the Food and Drug Administration at the time of presenting or negotiating the phase III protocol -- obviously, they're always allowed to propose, but you would say to the agency we have no particular preference whether you just do absolute change, percentage change, log transformation with an anti-log transformation of the results?

I guess to me I heard from all sides of the table and clinically also the percent change makes little sense. I must admit my own feeling about data of these sort are that numerical change is probably the best descriptor of what's going on in a patient and why get fancy when there's something straightforward there for you to utilize? I understand that sometimes these distributions may violate normality, but I think there are other tests for non-normal distributions.

So rather than transform or alter the data to make it easiest to use parametric statistics which give you the most power for a given amount of change in general, why

not have the underlying characteristics data drive what test you use and how to analyze it? I mean, that's what I was always taught in my simplistic way is not make the data so you can use your test of choice, but choose your test based on what the data shows.

DR. KILPATRICK: Dr. Stern, I want to, first of all, award you an honorary statistical degree.

(Laughter.)

DR. KILPATRICK: Secondly, I think you're going to hear three different points of view from the three statisticians.

I think this question is now redundant because my understanding of what is labeled as question number 1 is that we should augment the clinical evaluation by the numbers, by the counts, but not analyze them as a separate entity. That was my understanding. So I have no other suggestion.

DR. STERN: I guess my understanding is that you had to pass both the clinical evaluation and the counts.

Since two of the three old people are in disagreement, to refresh our memories, perhaps we should go around the room and clarify on that point what the opinion is. It's really not a vote.

DR. PLOTT: I thought we agreed on counts, just

on inflammatory or non-inflammatory, having a co-primary with the global.

DR. TEN HAVE: I think Dr. Kilpatrick is referring to the suggestion by Dr. O'Neal yesterday evening which is again this issue that I sort of threw out at the beginning of this discussion. I don't know where that's going to go. That's really probably an issue to be discussed for another day.

DR. STERN: I guess rather than getting into this discussion, just what your opinion is about it.

DR. TEN HAVE: My opinion. To me it's really a clinical decision, and if change in lesion counts is the most clinically meaningful outcome, then that's what I as a statistician need to hear.

If what happens in clinical practice is more complex, where the physician or the dermatologist first evaluates the patient on the basis of the size of the lesions, the color of the lesions, and then proceeds to a count, then maybe the research question should be more based on a combination of the lesion count and the global evaluation, which I believe Dr. Kilpatrick is advocating.

So as a statistician, all I can say is I agree with Dr. Stern's comment that you should stick with the data and try to transform it as little as possible and stay as close to the raw data as possible.

Of course, Dr. Berry had presented a number of different points of view, again implying that statisticians have a number of different points of view, as Dr. Kilpatrick indicated earlier.

DR. KILPATRICK: Sir, could you inform Dr. King as to what the question is because I would value his point of view.

DR. STERN: As would I.

There seems to be some ambiguity as to whether we were recommending that for a product to be approved, that it meet both the global test, once we have a good scale that really works, and the appropriate lesion count test, depending on whether it's for inflammatory, comedonal, or both.

As I understood Dr. Kilpatrick now, he was suggesting that if we had a good global scale that they didn't even have to make statistical significance. They were really a guideline perhaps for labeling or other things in terms of the magnitude of change, but if you had a good global scale, you only had to make it on that, and in a certain sense, the counts were irrelevant. That was not my recollection or that was not what I thought.

DR. TEN HAVE: The proposal I'm hearing from Dr. Kilpatrick -- but you should speak for yourself. But I think what I heard is you're talking about a combination of

the two, incorporating the lesion count and global evaluation together as a single outcome. So if both are clinically meaningful -- that is, if there's a situation where it's a non-inflammatory lesion but they're significant enough that you make a count, that's reasonable. Instead of counting 200, you're counting 20 or whatever in a particular area.

But I think what we were talking about last night is having a clinical indication that a count is needed. Once a clinical indication that a count is needed is made, then you do the lesion count. Then you'd have a lesion count outcome. But this implies, I believe, a more ordinal type outcome where you incorporate lesion count and global evaluation information on the same scale. I believe that's what Dr. Kilpatrick is referring to, not so much ignoring one or treating it as a secondary outcome.

DR. STERN: I'm sorry. I misunderstood you.

DR. TAN: Yes. That's exactly what I also referred to. You have a global scale. The global scale would account for the lesion counts as well.

DR. STERN: I think in operation that might not be simple, but I'd like to hear what Dr. King thinks about that and also Dr. Katz in terms of kind of combining numeric findings with qualitative findings based on scales that have some qualitative endpoint.

DR. KING: I guess I'm Darwinian in the sense of the agency already has data on studies or successes in the past, and the old thing, two out of three, and those kind of things that they've had. Yet, we're having this conference because there's -- I don't like the word "unhappiness" but there's uncertainty as to the efficacy, fairness, or reproducibility of the data.

I think the whole object here is we're trying to come up with some benchmark of success for is a product approvable based on the data you get. The question again is the question is "is." So it seems to me that the acneologists and so forth would say to you, well, there's such variability on global assessment of photography and so forth, there's such a difference individually that you could get way off. The purpose in the photographs is really to quantify the number of lesions there. You can transform it by computers. You can take out the redness and so forth and so on.

I don't think we're there. I don't think we have to continue the evolutionary. Can we get a better scale about if folks aren't clear, they aren't clear, but if they're near clear and the scale of two changes really says a significant effect, you're almost in the same category as a diabetic, which is if you take an oral agent and then exercise, you can move your blood sugar down from

180 to 120. Is that an improvement? Well, sure it is.

So I'm in favor of retaining the globals and improving the scale to include that not totally clear but that change and see what additional studies in the future would do using the new ways of measuring whether it's the old crazy glue or if it's a new photography that measures volume, anything. I think that we can't throw out the baby with the wash water. Where we are is built on studies in the past, and so if I were a sponsor, I wouldn't want somebody to invent a new scale and throw it out. I would like for all the studies that are in now to continue and to continue to use counts to verify objectivity.

Yet, clinically we know there's a difference between clinical significant and statistical significance. So I think the product that makes statistical significance but doesn't make a great deal of clinical significance will, in the marketplace, meet its maker, so to speak.

So I'd rather see the agency continue on here, try to bring in the academy for standardized photographs, and continue the numbers, and have the statisticians tell us are we clinically significant and statistically significant at the same time.

DR. STERN: Thank you.

Any more comments in that area?

DR. RAIMER: I just have a question about

globals on comedonal acne. I think it's going to be much more difficult to get a global assessment for comedonal acne, but is it unfair to require it for the others and not to require it for comedonal acne?

DR. STERN: Dr. Wilkin.

DR. WILKIN: Well, something I jotted down earlier from Dr. Stern was that if we look at the global for comedonal acne, the goal is to make sure it's not regressing, it's not losing anything.

DR. STERN: In my household, there's a big differential in impact between closed and open comedones. If you converted 100 closed comedones to 60 open comedones, that product at least according to the evaluators in my household would be at least 3 or 4 points up on a 6-point scale in terms of severity.

So we're now on question 5 I believe. Should lesion counts be assessed at multiple time points late in the study and averaged to increase power?

DR. KILPATRICK: I think I had spoken to that earlier in anticipation. Again, I'm very conscious that what we're doing is getting things very complicated even though we're trying to be simple because we now have the problems of inflammatory and non-inflammatory, and what I'm suggesting is that we have at least two time points other than baseline. So that means that we have to have

different procedures for these.

To speak specifically to this question, I don't think again they should be averaged to increase power. As I said, I think it's a matter of combining the evidence in the mind of the physician in some way because I don't want to get to spurious significance again.

DR. STERN: Dr. Tan.

DR. TAN: Yes. I think we discussed this yesterday. We want to have two time points so that we have consistent results. So I agree we cannot average them.

DR. STERN: I guess my own feeling is it's a little bit like the stock market. You only get to choose to sell the stock once and part of the protocol is you, given your agent, should pick a time that is going to help. If it works fast, based on the data, you pick assumed time. If it works slower but better, you pick a later time. And it's some other way that you have to commit yourself based on the information you have and the characteristics of the product of when you're hoping the market will be best in the sense of the improvement best. And that's the primary endpoint.

Certainly gathering information at other time points, just as we talked about other ways of getting out medically meaningful information that says not only at week 12 was it significantly better, that goes in the labeling,

but in fact, you can demonstrate in published studies that, gee, by week 4, 80 percent of the people were already improved. That's fine and helps information.

But I think in this kind of thing, again, keeping it simple and making people decide in advance what that single point is is the right way to do it.

DR. PLOTT: Today we pick a single point but we use an intent-to-treat analysis. So we take everyone that receives drug and we put all those patients into the analysis. Then we do a last observation carried forward so that for patients that are not responding, they're dropping out earlier, but we look at that last observation as though it were at the 12-week time point. In that way, we take into account better what the real situation would be in real life and make that estimation. I don't see how some of these repeated measures provide more insight into whether the drug works or not.

DR. TEN HAVE: Can I say something? There's sort of a debate going on in the statistical community about how dubious last observation carried forward is. There are some statisticians who don't think it's a big deal; there are some who do. I think it probably depends on the data.

I do know that the repeated measures analysis will give you a more accurate estimate and more accurate

test in general than last observation carried forward. There are simulations in the statistical literature that show, for most cases, that's the case. And you don't have to do any imputation.

The problem with last observation carried forward, it does involve an imputation and it involves a single imputation which gives you a false sense of lack of variability. And there's have been some effort in the statistical literature, especially the surveyed literature, to do multiple imputation to account for the fact that you are imputing something and that there's some variability in that prediction that you're making with last observation carried forward.

There are some arguments on the other side of the coin -- it's conservative -- in that if you project the worst outcome possible instead of the last observation carried forward, you make your estimates more conservative of treatment effect.

But I think from a statistical point of view, the repeated measures approach, the random effects approach, or generalized estimated equations preclude you from having to make an imputation which requires a lot of assumptions. And you still probably are increasing your power with the repeated measures analysis without having to do last observation carried forward. So I think there are

advantages actually that the industry doesn't see with the repeated measures approach that is there in terms of power. So I think we as academic statisticians have to do more of a --

DR. PLOTT: You're more familiar with the statistical. For us to do a placebo-controlled trial, it allows patients to drop out sooner, when there's a determination that this product is not effective, and still count that patient. I think that's good for patients because being stuck in a placebo-controlled trial and not being able to get out or the company losing --

DR. TEN HAVE: Right, but say you have a three-visit trial and they're still included in the analysis because you're incorporating those subjects at baseline and, say, visit 1. And they're still included in the analysis. They may not directly provide information at the third visit, but they're providing information in terms of the estimation of the standard error. They're providing degrees of freedom from a statistical point of view. They're providing information. It's just not as apparent as with the last observation carried forward. Nonetheless, you're not excluding them in the intent-to-treat analysis, and they're providing as much information as they're able to provide without having to do any imputation. So it's still ethical in the sense that the placebo patients who

are dropping out are still in the analysis.

But I think we have to do more educating I think.

DR. STERN: Finally, our last question, how should the efficacy outcomes of clinical trials be portrayed in labeling to be maximally useful to clinicians and patients? Specifically, what graphics and tables should be provided?

We saw some examples yesterday. Dr. Katz, would you start?

DR. KATZ: I don't know specifically what this refers to, but I would think that an efficacy outcome -- I don't think physicians look at labeling for efficacy. We prescribe drugs because we see presentations at meetings. We read articles in journals. We don't pick up a medication and say, well, let's see how I should use this drug. Rather, you might go to that after you prescribe the drug.

But I think to honestly present it to patients, contrary to what is done now where they might say, to quote from Dr. Bergfeld's study with the Ortho Tri-Cyclen, it would be said 60 percent of patients are improved with this medication, with Ortho Tri-Cyclen, frequently omitting the 40 percent placebo response. That's deceptive and it's repeated in abstracts all the time. So there should be

some requirement that it's 20 percent above placebo. That's quite different than 60 percent.

Or in studies quoted by Dr. Porres yesterday where it may receive statistical significance of a 52 percent response on an oral drug where 40 percent of placebo respond. Well, a patient should know over placebo response, 12 percent of patients, 1 out of 10, respond. How well? That's another story also.

To keep repeating that we have wonderful vehicles now, they make 40 percent of people better or 60 percent of people better. Well, the lactose tablet in the oral placebo trials with a 40 percent response or with the Ortho Tri-Cyclen with a 40 percent response, the other with a 42 percent response, nobody would propose that that lactose is effective. So I think it's important in the labeling to provide that information.

As far as what we're talking about today, I think it's important to put this has been shown to be effective in inflammatory or comedone acne or non-inflammatory acne. I think we should have that.

Another thing I would like to bring up. In these studies, all of which have a high placebo response, we're counting what we would all consider really not a very clinical significant difference of 2 papules, 2 inflammatory lesions. I mean, that's not very clinically

effective, though it may be statistically significant. We have to consider the lack of blinding in some of these studies, especially topical studies, where the irritancy certainly biases the investigator.

DR. STERN: I guess one of the things that I learned from Dr. Wilkin was -- 201-57(c) I believe it is -- that unfortunately there are some limitations on how we can present information in the package insert. What we discussed yesterday after the meeting was it seems to me what you really want is the equivalent of the abstract and one or two key tables or figures so that rather than having to go through everything, right up front you really get the abstract of these are the kind of patients we treated, this was the length of treatment, this is how many patients did so well, and here is a table that gives you the detail and the breakdown between drug and placebo all in one place right after the black box warning, if there is one, for that drug. That's what would make information useful, as I think Dr. Katz said.

In its current format, even for those of us who do a lot with drugs, the information gets to be -- you know, we want to see efficacy and safety together in adjacent sentences, and in the current format that may not be possible. So I guess one of our things is if the agency really wants to use the package insert for real

information, it has to think about ways of making some of it available in one place globally as sort of an abstract.

DR. BULL: Just as a point of information, it's important to note that the agency back in December of 2000 published for comment a draft proposed rule that will revise the professional labeling format. There were focus groups that were done with physicians and input provided to try to make this a more useful document.

I think it's important for you all to know that the agency sees the labeling as its primary communication of all the data. It is literally a scientific abstract of all of the data submitted to the agency for product approval. And it is our means of communicating to clinicians the assessment of the data and probably provides one of the few objective assessments of the data submitted for a drug to be placed into marketing.

So I would strongly encourage you to encourage your colleagues to engage with the labeling because there's a lot of effort put into composing the label. A lot of folks think that it's an industry document. It is done in collaboration with industry. Industry and FDA have to reach agreement on labeling. And I would encourage you to encourage your colleagues and yourselves to please read the label. It really represents a huge amount of work and effort on the part of our scientists.

DR. STERN: I guess I would argue as a clinician and as someone who has done a fair amount of editing that the label represents the equivalent of the whole article and really lacking is in fact the equivalent of the abstract. At least when I'm reviewing papers, I read a ton of abstracts. That's the hierarchy of many, many articles. Look at a fair number of tables and figures within articles for specific information to get some more detail. And the number of times in the literature that I, in fact, read front to back as opposed to going to sections that I've been alerted to by the abstract of particular interest, because of time, is just very limited.

So I think one thing, if you're communicating with people who are learning to get information in a certain way, presenting your information in a comparable kind of way is perhaps the best way to get people to pay attention to it.

DR. WILKIN: Actually the new labeling that Dr. Bull refers to has an abstract portion at the beginning of it.

DR. BULL: It's actually called a highlights section.

DR. WILKIN: A highlights section.

DR. BULL: We really are not able to comment at this point on specifics because it's being finalized, and I

wouldn't want to portray things that ultimately don't make it into the label. I think a lot of the concerns that you articulated will be addressed.

DR. STERN: Yes. I wasn't aware of any of those things, and certainly it sounds like it has the potential for really helping in information exchange.

DR. WILKIN: And then if I could just comment. The 201.57(c), that was the indications section of the labeling. It's 201.57 that is a very nice passage. It's multiple pages. It talks about all sections of labeling. (a) is description which is chemistry. (b) is clinical pharmacology. (c) is indications, so on. There's a very set format on how all of this needs to be laid out. I know I've gotten some comments from clinicians. Why do we put the dosage and administration section so far back where it's hard to find and things like that. But it's all laid out in a prespecified manner in the Code of Federal Regulations.

DR. SAWADA: I would certainly agree with why that is all the way in the back of that little tri-folded paper there. And adverse events also. This is what the patients are interested in when they look at the package insert rather than all the other stuff. Then for me when I look at that, I look at sizing, how does it come, what size do I prescribe.

MS. KNUDSON: I would just like to say that patients do read the package inserts. I would appreciate it very much if they were easier to read from a patient's point of view.

DR. STERN: Unless any committee members want to make closing comments, I would ask, Dr. Wilkin, if you believe we've at least made an attempt to address the questions that the agency has posed to us.

DR. WILKIN: I appreciate very much the comments that we've heard from the committee. I assure you that the FDA team, the statisticians and the clinicians, will be pouring over the transcripts of this important meeting. You've opened up the thinking to new indications and ways to approach those indications, and we will internally be working on a draft, and it will be something we hope to share with everyone in the future.

I think Dr. King may have made the point that we don't want to be held to future technology just yet. There may be some advances that will come and later we'll be able to figure those things out and move them in. These guidance documents are corrigible. You can always go in and correct things and update them and sort of thing. And FDA is always caught in this sort of time shift paradigm where we're approving the drugs that will be used tomorrow today based on the science of yesterday, which is now the

accepted science. So we're always trying to catch up and fabric that into it.

So with that in mind, I think we have a lot of good ideas to work on. As you know, it will be a draft guidance document. It will be open for comment. So there's still a lot of opportunity to get different views into this before there's any kind of final document.

Thank you.

DR. STERN: If no other committee members have anything they'd like to close with, I'd like to entertain a motion for adjournment.

DR. KING: So moved.

DR. KILPATRICK: Second.

DR. STERN: All those in favor?

(A show of hands.)

DR. STERN: I'd like to thank all the committee members for bearing with me and for their very active and helpful participation.

(Whereupon, at 10:45 a.m., the committee was adjourned.)