could consist of the Health Assessment Questionnaire, which you heard about this morning, and then also the Arthritis Impact Measure Scale, or AIMS.

It's important to say too that there should be no worsening on other quality of life measures in, for example, the SF-36.

The last claim is really the purpose for this afternoon's discussion and that is the prevention of structural damage. For this claim, the trial should be at least 1 year. In the past 3 years, we've seen a number of agents that have been evaluated for their effect for the structural claim, and this has generated -- I feel really this is not important to say at this point -- a lot of discussion both within the agency, outside of the agency, and this morning.

Anyway, the guidance actually describes some examples of outcome measures for this indication. The first they discuss is slowing of x-ray progression, and this would be done by a comparison from baseline to the 54-week or week 102 or even longer, using a Larsen or a modified Sharp score.

The second example for the prevention claim would be a prevention of new x-rays. The guidance document here just simply describes the landmark comparison of progressors and nonprogressors.

Then it also leads into some discussion of other measurement tools. It describes, for example, the use of the MRI, and we touched upon that this morning. The document describes that some of the extrapolation for the interpretation of radiographic change or lack thereof to patient benefit remains undetermined.

However, regardless of the products that are developed to show a delay or prevention of structural damage, they have to be shown to have a clinical benefit, either first or even under accelerated approval. The guidance document discusses the development of agents, and I'm going to quote it: "not intended to affect acute inflammation, but are designed to prevent or slow joint destruction by other means."

Therefore, the first indication for such a product would be that they would be for slowing for radiographic progression as a surrogate marker, but you would need to show clinical benefit either later in the trial or in a separate trial. I should point out the document is hazy. In fact, it doesn't even define what is meant by clinical benefit.

So, needless to say, in the background of what we heard this morning and what I sort of inferred through my background presentation, there are a number of considerations regarding the document in light of these new

products that we've been seeing.

There has been continued discussion regarding the most relevant outcome measure for signs and symptoms. For example, should an ACR-N, whatever that would be, be more useful than just the ACR20 alone? But such a comparison may result in a statistical difference between treatment groups where neither achieves an ACR that would be greater than an ACR20. And would this, then, be suggestive of clinical benefit? In other words, would we accept an ACR15 versus an ACR5?

Another consideration is should some of the individual ACR components be used instead of just the ACR20? This is what is being accepted over in the European Union where they accept tender and swollen joint counts and HAQ in support of their claim for signs and symptoms.

The pros and cons of landmark and response-over-time analyses continue to be discussed. For example, a product with early onset of activity may achieve a greater area under the curve, even though both study agents, including the placebo or the control agent, may achieve similar degrees of effect by the end of the trial.

One of the major points of discussion regarding the topic of disability or the function is the effect of missing data on these analyses, predominantly due to dropouts where the number of patients who dropped out tend

to increase with time because of lack of efficacy. As you recall, the guidance document is asking for long-term trials for this indication. So, this leads into the discussion of the length of trials in light of the relatively early effects upon functional outcome measures that we are seeing soon after initiation of these newer modalities.

Then last, but certainly not least, are the considerations regarding structural damage. I think the one that is most prevalent among all of us at the agency has been the use of the word "prevention" because, as has been pointed out to us, to many people the use of the word "prevention" implies that the patient population being studied are disease-free. As you heard, I don't think any of the products that we've seen recently have been in a disease-free population.

Now, the document does discuss prevention of erosions, but it's very vague. It just says a landmark analysis between progressors and nonprogressors. So, what do we mean by progressors and nonprogressors? This has obviously come up this morning. Do we use the smallest detectable difference or anything greater than 0? How do we define this or should we define this? What role do the more sensitive imaging agents or modalities have with the assessment of new erosions in patients with early disease,

and how does it affect patients with other stages of the disease?

In addition, with the clinical development of these newer agents, we realize that the guidance actually fails to discuss other structural outcome measures, for example, the reduction of erosions, the healing of erosions. Again, how do we measure for these effects? You can't just say to do it, but give guidance on how to do it.

Lastly, in patients who do not experience clinical benefit, is the delay in radiographic progression of their disease that was seen in a subset of patients that we discussed this morning a surrogate, and if so, what does that surrogate mean for clinical benefit? And then the natural extension, as was touched upon this morning, is how do we define it, how do we study it?

So, in summary, actually the preceding 5 years have been very exciting in the development of therapeutics for the treatment of patients with rheumatoid arthritis. These modalities provide clinical benefit to a great proportion of patients. But in addition -- and that's the point of the talk this afternoon -- these products have challenged all of us to continue to discuss how best to measure the clinical effect and allow further development of these new agents.

Thank you.

DR. SCHWIETERMAN: I thought we'd go right to Dr. Strand's talk, just to give a brief overview, and then we can have a discussion, if that's okay.

DR. SIMON: Fine. We're calling up Dr. Vibeke Strand, who is a biopharmaceutical consultant and a Clinical Associate Professor at Stanford University.

DR. STRAND: Thank you very much, panel, for the invitation, and ladies and gentlemen.

I have to tell you that unfortunately a virus blew up my computer last night. I couldn't open any of my files. So, I have to show you some traditional old slides.

So, what I wanted to do today and actually what I've been asked to do today is to review the data with the recently approved and presumably soon-to-be-reviewed products that have radiographic outcomes. These include, of course, several different products: the leflunomide trials, which actually looked at methotrexate and sulfasalazine and compared these traditional DMARDs to placebo, as well as the biologic agents, infliximab that you heard about this morning in great detail, etanercept, and anakinra, which presumably will be reviewed sometime in the fairly near future.

Now, I am taking the data from published references, and I do have a reference list, which I'd be glad to share at the end of the meeting. The slides on

data that have not yet been published but presented and have been presented in abstract form were kindly provided to me by both sponsors, as they're used for their speakers bureaus. What I basically want to do is to review these data sets but not, in fact, to compare them.

So, if we start first with the leflunomide data set, I think it's an important data set because it also tells us that the gold standard DMARD, methotrexate, does in fact slow or delay radiographic progression, and that sulfasalazine has a similar effect over the short term. These are both in comparison to placebo.

The Sharp scoring method was utilized, and it's very similar for all of the different trials that I'll be reviewing.

X-rays were done at baseline and endpoint in all of the trials. The films were read by Dr. Sharp using the modified Sharp method and by Dr. Larsen using his method. Those results did correlate, even though the Larsen scores predominantly score only erosions.

There was a formal 12-month intent-to-treat analysis in the 301US or ULTRA trial, and this was because patients may have entered into alternate therapy or have exited for active therapy, but they were recalled at 12 months for x-rays.

The clinical data was comparable in patients

both with and without x-rays at endpoint, and a variety of sensitivity analyses were performed to account for the missing data, and that has in fact been published.

Now, the demographics of these protocol populations are interesting because they probably do account for some of the differences in the radiographic findings. Two of the studies were placebo-controlled, and thus we have placebo to look at and understand the methotrexate effects, as well as leflunomide and sulfasalazine. The doses of methotrexate in the U.S. and the European study were fairly comparable, although the median dose was higher in the U.S. study. The sulfasalazine dose was a traditional dose.

The U.S. study was a 12-month placebo-controlled trial. The 301MN study was a 6-month controlled trial, after which placebo was allowed to exit for active treatment, and blinded treatment was continued for a total of 2 years. The 302MN was an active-controlled trial, leflunomide versus methotrexate, that was continued also blinded for 2 years.

Now, although the disease duration looks fairly similar in these two protocols, it's significantly less in 302MN at 3.7 years. In fact, a significant portion of these patients in all three protocols have less than or equal to 2 years of disease or are DMARD-naive, on the

order of about 33 to 50 percent. This protocol had about 40 percent of patients with early disease and about 40 percent of patients with more than 5 years of disease, a bimodal population of DMARD- or methotrexate-naive patients in the United States, whereas this was rather evenly distributed and this protocol had an excess of patients with early but very aggressive disease.

The mean HAQ disability indices were different across the protocols: 1.3 in the U.S., 1.7 to 1.9 in the sulfasalazine controlled trial, and 1.5 in the methotrexate active control.

Now, as a benchmark only but as a means of understanding what progression might have occurred just during the 12-month period of time of the protocol, an estimate or a predicted yearly progression was derived by taking the total Sharp score at baseline and dividing it by the mean disease duration. We've come up with the estimates of around 3.3 to 3.7 total Sharp units in 12 months in the U.S. study; 5.7 to 8.1 in the MN301 study, but over 6 months, this would be half that amount; and about 6.5 to 6.7 in 302MN.

To look at this benchmark comparison and not on a statistical basis, we can ask whether this would be a relative estimate that would have any accuracy. If you look in 301MN at the 6-month data, the predicted

progression is in gray at 4.1 points, total Sharp score, in 6 months, and the actual progression in placebo was 5.9. So, if anything, this estimate was a bit of an underestimate but a reasonable benchmark.

If you look in the U.S. study with the placebo, the placebo actually did not progress as much as might have been predicted, but we know that 63 percent of these patients had had active therapy at the 12-month endpoint, whereby about half of them had received alternate therapy for a mean of about 6 months and the other half had received active therapy, having left the protocol prior to month 4 due to lack of efficacy.

Now, despite that, the statistical comparison against placebo here in this U.S. study was highly statistically significant for both methotrexate and for leflunomide, so that even though these patients may have had as much as 6 to 8 months of active therapy, a full 12 months of therapy made a very significant difference.

Again, if you look at the placebo-controlled trial in the European 301MN, you can see that even with 6-month data there's a highly statistically significant difference with both leflunomide and sulfasalazine versus placebo in total Sharp scores.

In the 302MN study, these two numbers are equivalent between methotrexate and leflunomide.

Very quickly, just to look at the 6-month data with the 301MN study, you can see in the active therapy groups when treatment was continued over a full 12 months, these results were maintained and remained statistically significant even against the placebo at 6 months.

Now, if we talk about the percent or number of patients with no newly eroded joints as some type of an idea of what that could be defined as, no new erosions, we can see that there's a high percentage of patients, in fact, in all of the treatment groups who do not have newly eroded joints, and there's a fairly high percentage of patients in the placebo groups that have no newly eroded joints, indicating that there are a large number of patients in these skewed populations who have no progression, at least by erosions, in terms of involving new joints.

If you look at the mean changes in the erosion and joint space narrowing scores, one can see in the U.S. study that in fact the medians are also 0 for both the placebo, as well as the active therapies, and the mean changes are quite low but involve both erosions and joint space narrowing and are significantly less than the active treatment groups than in placebo, but indicating, in fact, progression in both groups, as measured by both joint space narrowing and erosions. Yet approximately 50 percent of

patients will have had scores that are 0 or negative.

If we look at the 301MN study, we see some similar differences although the placebo is obviously quite higher. It's on the same scale of 0 to 2, but you can see now that there is a median increase of 3 with the placebo and 0.5 with sulfasalazine. However, in terms of the erosions, there is still a large percentage of patients who do not have erosions.

Now, the means in all of these studies, including the next one, have standard deviations that exceed the actual mean and the ranges range from minus, a negative score, but at least two digits to a positive score of as high as two digits. So, we're looking at skewed populations with broad ranges of results.

Here we see in this study with methotrexate and leflunomide that, in fact, the medians are again 0, and the erosion and joint space narrowing score progression over 1 year are very similar and they're statistically equivalent.

In fact, there is a correlation with response in the ACR response in the U.S. study, but in fact there's a negative correlation in the 301MN study where all active treatment groups actually look better in the nonresponders, a minus 1.7 and a minus .4 score for leflunomide and sulfasalazine in the nonresponders versus a 1.3 and 2.7 in the ACR responders. In a similar fashion but opposite

direction, there actually does look to to be a correlation between ACR response, although mild, because the scores are lower in the responders in the U.S. study in the two active treatment groups than in the nonresponders.

In terms of coefficients of correlation, looking across all three studies, CRP was associated with response in the U.S. study but to a very low degree, .17, which was statistically significant. There was a better correlation of .22 in the 301MN study, and there was a correlation of .15 in the 302MN study. These were statistically significant although very low. Only the AUC of ACR response and the ACR20 response correlated with the US301 data in the active treatment groups, as this data shows you.

If we now move on to the ERA study, we're looking at a very interesting population of early disease, and we're looking at an active controlled trial, looking at etanercept in two doses versus methotrexate. The scoring method was a modified Sharp scoring method again. X-rays were taken at baseline, 6 and 12 months. The feet were included as they were in the previous analyses, and they also looked at the Rafingen method by Rau. In this situation, there were two of six readers who read the films, and the inter-reader variability was shown to be .85 as a correlation coefficient. The sequences of films were

blinded as in the other study as well.

It's a different demographic population. Again, it's very important to understand these patient populations because the x-ray data looks different accordingly. Mean age is low but patients have all got disease duration of approximately 1 year, and most of them are either rheumatoid factor positive or have erosions. They were included in the protocol on that basis. The baseline HAQ disability indices were 1.4 to 1.5, and prior DMARDs ranged between .5 and .6. There were similar uses of nonsteroidals and steroids.

We can use as predicted yearly progression again only as a benchmark, but if we see it, it's 9.5 in the methotrexate group and the progression in the methotrexate group is 1.3.

In the etanercept 10 milligram group, it's 1.4 versus a predicted progression of 8.3, and in the high dose group, it's 0.8 versus a predicted progression of 8.7. These are statistically not different in the active treatment groups.

When we break it down, we see that there is more of an effect on erosions in the etanercept high dose group, and by the way, these are again mean scores and the ranges are from negative numbers to positive numbers. The standard deviations exceed the mean scores and the medians

in this particular 12-month data set are also 0 indicating that approximately 50 percent of patients do not appear to progress in terms of increasing a score above 0.

If we look at the number of patients with no newly eroded joints at 1 year, we can see that there's a fairly high group of people and a higher percentage in the etanercept high dose. Interestingly enough, these are very well represented by the number of patients who have no erosions at baseline. In other words, those patients who have erosion scores of 0 at baseline will often not develop new erosions over a 12-month period in a protocol as we've so far observed.

By the way, the clinical correlations in terms of response by x-ray and AUC of ACR-N was low but predictable, again about a 0.15. The AUC of the CRP was the best correlation with a coefficient correlation of 0.45 in this data set.

If we now look at ATTRACT, I'm just going to review this very quickly simply to point out the comparabilities. Again in terms of the methodologies, slightly different numbers of joints being scored, and the feet were scored from 0 to 10. There were two readers, as you had heard, baseline and 6- and 12-month data. Again, there was a large percentage of patients who had final films, and a sensitivity analysis was performed to account

for missing data. Again, the clinical data in the patients with and without x-rays were very comparable.

You've seen these demographics before. This is a patient population who has failed methotrexate. They have aggressive disease, fairly long duration, 8.4 years median duration, and mean duration of 10.4 years. They had failed on average as a median 3 DMARDs, and many of them had previous joint surgery. And there are high baseline HAQ scores of 1.7 to 1.8.

If we look at the comparison here again, this is the mean change in total Sharp score at week 54 by a mean of 7.0 in the placebo plus methotrexate group versus very significantly less and statistically significantly less progression in all of the active treatment groups. The estimated yearly progression was approximately 7.4, although it hasn't been actively calculated.

If we look at the median, you can see that now we can get to a 0.0 for the entire infliximab group population, and the control group, the methotrexate failures with placebo, have a median of 4.

Again, there is a disconnect, as you may want to call it, between clinical responders and nonresponders, or ACR20 responders and nonresponders, and those who have progression by x-ray, although there is a suggestion here that there is better response in the responders.

Now, if we move to another patient population -- in fact, this study was the first study with radiographic data that was placebo-controlled. It was 6-month data and it was originally performed using Larsen scores. These have now subsequently been reread in blinded fashion using the Genant modification of the Sharp scoring method. In fact, the feet were not included in these films. So, they had hand films scoring 28 joints for erosions, 26 for joint space narrowing, a slightly different grading for it and a summed total score. The number of joints are less because Genant believes that this modification allows it to be less error due to deformity overlying shadows when you place the hands on an x-ray film.

X-rays were done again at baseline, 24 weeks, and 48 weeks. The placebo patients at 24 weeks were then allowed to be re-randomized to active therapy on an open label basis. However, the active treatments were continued for a full 12 months blinded treatment. These were again scored in pairs or triplicates, and the sequence of films were again blinded.

This is a patient population with a mean disease duration of somewhere between 3.5 and 4 years. Many of them were rheumatoid factor positive. About a third of them had erosions at baseline. Interestingly,

there are as many as 25 to 30 percent of them who were also DMARD-naive in this population. Concomitant steroids were a little bit lower than in the U.S. protocols that I just reviewed, and concomitant nonsteroidals were very consistent. The baseline Sharp scores, using the Genant modification, ranged between about 25 and 30.

If one now looks at the estimated yearly progression, again in gray bars, just as a benchmark, we can see that the placebo progression over a 6-month period of time was just about equivalent to what the estimated progression might have been. In each of the treatment groups, 30, 75, and 150 milligrams, there is statistically significantly less progression than in placebo.

If we now look at the continuation data whereby the placebo group here is now continued on to active therapy and they are all summed, we can see that the first 24-week therapy in orange in each of the treatment groups is actually less in the subsequent 24 weeks, with the exception of the 30 milligram dose group. So, in other words, there's more effect over time in these treatment groups. Clearly the placebo has a very significant change between the first 24 weeks and the second 24 weeks with active therapy.

Looking at it from a statistical point of view where the placebo patients are then dropped out after the

first 24 weeks, we can see that in the initial response, the majority of the response may well be in joint space narrowing, but over the second 6 months of active treatment, there is also a very significant effect on erosions such that the total scores are considerably improved over time.

This is looking at all of the dose groups merged with a total erosion score of 1.2 and a joint space narrowing of 0.6 over the first 6 months and subsequent 0.6 and 0.6 over the second 6 months, indicating again that the early effect is more predominant on joint space narrowing which is maintained, but then erosions are also affected over the second 6 months of treatment.

If we look at the patients who now have scores of 0 in erosions, it's 42 percent for placebo, 53 percent for all active, and this is at the 6-month time point; and 44 percent versus 59 percent in terms of 0 scores in joint space narrowing at 6 months, for a total of 33 percent of patients who have a score of 0 in placebo versus all the IL-1ra patients, 43 percent, with scores of 0 at 6 months.

So, what we have, in fact, is a group of studies that have looked at basically methotrexate and leflunomide, sulfasalazine, placebo, placebo superimposed on methotrexate failures, and placebo here again. And in data not shown, we're basically seeing that the 6-month

information both with the MN301 study and the IL-1ra study, but also with ATTRACT and with the ERA study, that very significant effects are evident over 6 months' treatment in radiographic outcome even though 12 months has been the benchmark selected.

We can also look that each protocol population is different in terms of disease duration and baseline score, and one can use these estimated yearly progressions only as a benchmark, but in fact they may be a fairly reasonable benchmark of what might have been expected in a group of patients just during that protocol period of time who remain untreated. So, in other words, we're looking at an estimate of 4.1 which was exceeded by 5.9 in placebo and MN301, an estimated yearly progression of about 7.4 to 8 in ATTRACT, which was just about met by the methotrexate failure patients receiving placebo at 7, and the European IL-1ra data where 6-month progression was expected to be 3.6 and was shown to be 3.5 in the placebo group.

What we also see is that any way you want to modify the Sharp analyses, you can score them as a total score of 422, 440, 348, or even 202. We are seeing statistically significant differences between active therapies and controls, that we have a series of data sets where sensitivity analyses have nicely accounted for missing data. And we have one reader in the ULTRA and

MN301 and 302 trials, which was confirmed by a second analysis of Larsen scores, and multiple readers in each of the other studies.

We can conclude that leflunomide and methotrexate are effective against placebo and that they appear to be equivalent to sulfasalazine over a 6- and 12-month period of time; that infliximab, as was discussed this morning, is effective in the doses and dose schedules used in the ATTRACT trial against placebo superimposed in patients failing methotrexate; that etanercept is effective and was statistically equivalent to methotrexate in a patient population with very early RA; and that anakinra appears to be effective against placebo with continued effect over a full 12 months of active treatment.

What we've learned is that each protocol population is unique. Their baseline demographics in part may determine what their baseline Sharp scores are. Their rates of progression appear to be different, and it's not very easy to predict what the outcome will be prospectively before the patient population has been enrolled.

The estimated yearly progression may be a reasonable benchmark to understand the data. It should, of course, not be used for any statistical comparison.

We have a variety of modifications of the Sharp analysis, including Sharp's modification of Sharp himself,

and I think he continues to keep reinventing himself, I hope. But on that basis, we see that the scores can be scored in different ways, that the feet can be scored on scales of 0 to 5 or 0 to 10. Is it more important to have more joints to assess or is it better to have fewer joints? A 28-joint count may be more sensitive and less variable to change than the 66-68 traditional joint count. Would we find the same thing here with these various modifications?

I think the important thing really is that we see that there is marked variability in terms of change. These are very skewed populations of patients. It may be appropriate to use a median, or it may be more appropriate to use a mean, as one cannot differentiate from active and placebo with median scores.

We know that a large majority of patients in any of these protocols receiving placebo do not progress by a variety of definitions that have been used to say no progression. It's really still up in the air I think as to whether we can learn more about how to treat skewed data, in fact, whether we can even apply Bland/Altman to skewed data to try to understand SDDs. I think what we've heard so far about SDDs is that they, in fact, exceed the active treatment total scores in each of the protocols that they've been applied to.

Finally, should we be expressing the data in

terms of the total Sharp score and as well looking at erosions and joint space narrowing, or should we be able to say that since they are probably biologically separate processes, that we should be able to dissociate these scores and look for changes and hope to consider that benefit to patients could be associated with significant change in one or the other and not both?

Can we define healing? Can we actually even define what no progression is? We've seen several different definitions. We've seen several different definitions of no newly eroded joints. And what does that mean in the sense if we're looking at multiple joints? Some may heal and some may actually worsen. How do we actually look at a definition of healing if we're reading all of the films blinded to sequence?

The correlations between ACR responses, HAQ scores, CRPs, sed rates, even AUC analyses of outcome are actually very low with radiologic responses, suggesting that we may be looking again at different processes that are all part of a very heterogeneous disease that we call rheumatoid arthritis.

Finally, we still have to learn a lot about the statistical methodology that we're using. What is clinically meaningful? What is statistically significant? Can we take changes in group populations and apply them to

individual patients?

Are these sensitivity analyses the way to deal with missing data, as patients will inevitably not comply completely with the protocol, and we will always have at least some x-rays that are missing?

And what do we do about variable assessment due to multiple readers? Because the SDD is probably not just technique dependent and inter-reader variability dependent and protocol population dependent, but in fact, it reflects that we are studying a very, very heterogeneous disease.

So, I thank you for your time and trouble in listening to this rather dry set of data.

DR. SIMON: We'd like to thank you as well.

Are there any questions for Dr. Strand?

It does raise the issue, doesn't it, of are these different patient populations? If you have a group of patients that have no baseline erosions within the time frame that we're talking about and those patients don't also accrue new erosions without therapy, is that the same patient population as a group of people that erode in the time frame of the study? Is it fair to compare between those two patient populations?

Furthermore, is a new erosion that one sees because of the technique that one used, a flat versus a 3D surface -- are you actually not seeing that erosion or has

there been some other change taking place just in the way the x-ray was done? One can figure out any number of different circumstances.

And if in the area of interest, you're looking and see no new erosions but an area that you don't study has erosions, does that mean the disease is not progressive because you just haven't looked where they might have new erosions?

Bill?

DR. STRAND: I'm glad that wasn't a question and I can sit down now. Is that right?

DR. SIMON: Absolutely.

DR. SIEGEL: After each speaker, I keep thinking we have too many questions to answer and it's only getting worse.

(Laughter.)

DR. SCHWIETERMAN: Well, Dr. Simon, I think we actually had a very nice discussion this morning, and I think without further ado, we might as well just plunge right into the questions.

DR. SIMON: Well, in your packets is a second series of questions. Similar to this morning, they have paragraph prefaces, and I think that they really were precipitated somewhat -- although they were already written before we even got here, but they really were precipitated

by the discussion we had this morning.

So, you've already heard about the agency's guidance document for sponsors developing therapeutic agents, and it lists a claim for prevention of structural damage. The requirements are study of at least 1 year in duration. There have been two agents recently approved by the agency showing effects on radiographic progression in patients treated for 1 year. They were indicated for retarding or delaying structural damage and not preventing structural damage, and that was because in their patient data set, many patients -- they don't say how many -- were observed to have worsened structural damage on treatment. In this context, a prevention claim seemed inappropriate. Not too dissimilar from the discussion this morning.

So, the first question is, is prevention of structural damage a viable claim in rheumatoid arthritis given that, even following treatment with very active agents, some patients are likely to have some evidence of some disease progression?

So, for example, this morning we saw data that perhaps 6 percent or so still had ongoing evidence of damage. Is that enough? Does it have to be 20 percent, 50 percent? What would the committee like to think about in that regard?

DR. ELASHOFF: Is structural damage being

defined as these Sharp scores for radiographic or does it have some other definition?

DR. SIMON: Right now, as we understand it, according to the guidance document, as defined in the guidance document, this form of structural damage, which presumably is radiographic measurement damage, would be measured by the Larsen or modified Sharp technique.

DR. ELASHOFF: So, that's what we're to take the definition as.

DR. SCHWEITZER: I'd like to say two things. First of all, I would again try to bring the discussion against limiting it just to x-rays as a way of measuring surrogate damage. I would just say some kind of anatomic imaging to measure structural damage.

Secondly, I said before since rheumatoid is probably a somewhat protean disease and there's probably maybe more than one or subgroups of patients within a rheumatoid population, as long as I think there is a majority -- and we use the terminology "prevents progression" rather than "prevents structural damage" I think as long as it's a majority.

DR. FIRESTEIN: Again, as was discussed this morning, there's virtually no drug that has 100 percent efficacy in 100 percent of the patients, and the bar would be too high in that sense. So, again, a reasonable

approach is similar to what you just described, and that is that if in toto the group has delayed or prevented progression, depending on what you wanted to call it, then that's appropriate. But I think it would be inappropriate to try to ask for 100 percent of patients or even 99 percent of patients to have complete arrest.

DR. SIMON: It's amazing because in the context of signs and symptoms, we've agonized over some form of composite scoring to give us a sense of objectivity, ACR20, 50, 70. Would we have to figure out some kind of similar format for composite scoring? Would we draw a line somewhere where if you only had 40 percent, that wouldn't be enough, and if it was 60 or 70 percent, it would be enough? How do feel about that?

DR. FIRESTEIN: Well, ultimately I suppose we will have to draw some sort of arbitrary line, and I don't know what the number is.

DR. SCHWIETERMAN: Lee, I think that's an excellent question. To put this into perspective, the point of the agency guidance document was manifold, but the most important one of which was to characterize these agents over the long term. To set arbitrary thresholds for particular proportions or particular numbers and so forth, especially when the clinical relevance of those is unknown, sort of just goes against the agency's grain, so that if

you're at 40 percent, you get the claim, if you're at 39, you don't or whatever it is.

That having been said, I think that the guidance document has actually helped the field a great deal because now we have long-term data on these particular claims. So, at the risk of opening a can of worms -- perhaps it's open already -- what we really want in these documents -- and this has come up in internal discussions -- is a characterization of these agents so that physicians and patients have some way of assessing their likelihood of affecting multiple different outcomes without their being unnecessarily arbitrary or misused words that are out there.

I don't have any particular suggestions as to what this might involve, but this is in itself the very problem, if you will, with the guidance document as written because we have claims that are somewhat artificial out there, or at least the words are very difficult. It was very easy to concoct these words in 1995 and 1996 because we didn't have anything, but now that we're here, people want to make claims about them.

Just putting it all out on the table, does the committee feel that there needs to be reconsideration perhaps of walking down this radiographic outcome? We could talk about prevention. Then we could talk about

reduction. Then we can talk about the proportion that have reduced by a certain amount. Then we can talk about the complete elimination of erosions and so forth, none of which we can readily relate to clinical outcome measures, only to our sense of what that might or might not mean.

So, I think that this is what the problem has been in our discussions here, and this is frankly what the problem has been with our sponsors, is that they've acted in good faith with this, yet at the same time we owe it to the public to give accurate information. Does anybody have any comments on that?

DR. SIMON: I do.

(Laughter.)

DR. SIMON: What a surprise.

I think that that is exactly the problem. I think that putting the data in the data section of the label that's referable to that particular claim is very critical and it needs to be all the data. I think that that evolution has been very important.

The dilemma, of course, one could ask, well, why make any claims? Just show the data and let the individual decide what to do. That, of course, doesn't work in this world. So, we have to do something.

Unfortunately, still we're using inadequate techniques to be able to answer this question.

But I do think that the committee probably would believe -- and please join in -- that we need to relook at the question. We need to rewrite the document that reflects what we now know. To be able to distinguish among words would be very important. We probably have more possibilities than we had before that probably need to be included.

We probably need to be very clear that if we prevent, what are the various different thesaurus words that would be other terms that one could use for prevent. I think we also have to raise the question that in this real world of nuance, in differentiating among products that look very similar, the word is very important in what the claim will mean. Unless we're willing to distinguish among products that have very similar effects, which I'm not sure we're willing to do, then we have to be very careful about the words that we apply.

Does anybody have an issue with that?

(No response.)

DR. SIMON: Outstanding. Consensus.

Dr. van der Heijde

DR. VAN DER HEIJDE: I think an extra issue I want to bring up is that we really have to look within the trial because we are using randomized, controlled trials exactly to half the control group. So, if you are looking

at only at percentage of patients nonprogressing or progressing, whatever you want to use, and then you're comparing that across trials, that might be very dangerous because a control group is completely different. It has been shown by Dr. Strand that in some trials, even in the placebo group, the majority of patients were not progressing, while in other trials, the majority of the control group is progressing. So, if you have a reduction in that type of trial, it's much stronger than in another trial. I think that's important to keep in mind.

DR. SIMON: Do you have a proposal on how to go about doing that?

DR. VAN DER HEIJDE: Well, if we reiterate the OMERACT discussions that have been going on in May last year, then we decided as a group that the primary analysis should be on the group level. So, that's the first thing. You have to compare the two groups, and only if there's a statistically significant difference, then you go further for secondary analysis.

If we look 10 years back, for example, in '89 I published a trial comparing sulfasalazine and hydroxychloroquine, and we looked only at group levels. Then people accepted it with skepticism. Now we are only 10 years later and there has been such a big change. Now we are saying, oh, yes, we know that we can reduce it, but

we cannot in all patients. So, there's really a change of expectation.

So, I think the first thing you need to do is the primary analysis on the group level. If you find a statistically significant difference, then you want to see what does this mean on an individual patient level. How many patients really benefit from this therapy? Then you can do it in several ways because you need to use a cutoff point. That's a separate issue, how to define that.

But then you could apply the number needed to treat concept. That can be used and it's also used for other drugs. So, how many patients do you need to treat to have the benefit for one patient? That takes into account what's happening in the control group because then you are looking for the risk reduction between the control group and the active treatment group. So, you're taking that into account and you can calculate a number needed to treat. I think that could be very helpful comparing across trials.

DR. SIMON: First Carl and then Dr. Emery, please.

DR. WINALSKI: One question. I'm undecided of how things should be looked at, but it seems to me that with the signs and symptoms, they decide either they're a responder or nonresponder. Why should radiographic

progression be looked at as a group rather than as a responder or nonresponder, as you do with the SDD measurements? And the SDD measurements also help take into account how fuzzy your ruler is, if you will, with those readers. But I can see one problem right away, which is it raises the bar pretty high.

DR. SIMON: Dr. Emery?

DR. EMERY: Thank you.

I just wanted to introduce another issue if you are looking forward, which is the one that we've come up against both in our OMERACT group and MRI group. If you are actually going to look at a comparator, which is actually active disease, you have a great deal of problem if you're going to look at 2-year data. Already the ethical issue is such that we can't do studies anymore because the correlates with CRP, for example, are felt to be too good to leave patients with an active CRP. It's unethical for patients and we're no longer allowed to do it.

If you're looking forward, the only way you can show differences with x-ray, because of the SDD being so large, is between active and inactive therapies. If you're going to ask for 2-year data, you're making it impossible to get because patients drop out, get steroids, and produce some of the confounding data that you've just seen.

So, the only way you're probably going to be able to do it, we're now studying for the first time longitudinal data comparing all modalities. Those studies are there. We are analyzing them. You do have to use the more sensitive techniques because they're the only ones that have the power to show a difference between active therapies. I don't think there's going to be a great deal of value in devising criteria that are just going to be impossible to reach.

We've already reached the stage you can't leave patients with active disease. We have to put in our consent that patients who get placebo will be out of it if they're not improved within the time course of the half-life of that drug because otherwise it's felt unethical. Americans still seem to be able to do placebo studies. We can't. I think you need to think very carefully if you want patients to go into studies, that you're not harming them by leaving them untreated because the data are there to show that they are harmed.

DR. SIEGEL: Aside from where the bar is set, I would agree very strongly with Dr. van der Heijde that there is really good reason not to stray too far from looking at aggregate rather than at number of patient responder data. There's actually a number of reasons.

One is you can define a response that's

meaningful on an individual basis, an ACR20 or an SDD, based on what you think is a real thing that isn't just chance fluctuation or variation, but smaller responses, if they're seen consistently across large numbers of patients and statistically significant, are real. They may have less clinical meaning, but you don't have to question whether they're real. If you treated 100 patients in a trial and every one of them showed a 10 percent reduction in their joints, you wouldn't have any ACR20's, but you'd be sure you had a drug that had an effect if in the control arm none of them did is what I'm saying.

So, you lose certain information. Even if you don't set the bar that high, just by drawing cut points you lose a lot of information, just number of responders, number not responding. That's part of the problem with the whole concept of prevention to the extent that we think of it as an absolute, that we're starting to think of the proportion who do progress and who don't progress. What you have in this trial, what you have in other data sets are two treatment arms where there's a distribution in each arm and they're different from each other.

That gets to the issue I think that Dr. Firestein raised. If we set the numbers too high, since no drug works in everybody, it becomes impossible. But what's at the table is not setting a bar so high that it's

impossible to make a claim, but more what's the appropriate nature of the claims that can be made with data, such as much of the data that Dr. Strand summarized.

DR. SIMON: Bill, did you have another comment?

DR. SCHWIETERMAN: No. I just wanted to make a brief comment about the U.S. mandating placebo-controlled trials. I think the points were all taken about the need to have equipoise in trials and not unnecessarily mandate. I'm not sure the implication was that we mandated long-term placebo-controlled studies. We certainly don't. Rather, we look to the patient population, the standard of care, and so forth.

With that having been said, however, I think that there is a role for placebo in many of these trials, and if properly designed, you can do that.

DR. SIMON: Dr. Sharp.

DR. SHARP: I want to agree with a lot of what has been said, but I want to get back to the placebo issue first.

I've been opposed to placebos for 10 years now because I've been convinced that we have drugs that do something. I agreed to participate in the leflunomide trials because of the escape clause for the placebo-treated patients. I'm not sure that we can even countenance a 4-month placebo with escape today now that we have more

effective treatments. I don't think we'll get it by institutional review boards in this country.

The ATTRACT trial has been called a placebo trial. The placebo here was a blinding mechanism in a trial of combined therapy versus monotherapy to make it possible to blind it. It wasn't a placebo in that the patients weren't getting active treatment.

I think one of the main issues that face us in developing new drugs is how are we going to test them without testing them against a placebo. There are several issues involved. We've seen two trials in the last few months that have come here, and they've been designed entirely differently. I think we probably will see additional trials that are designed differently, and perhaps over a period of the next few months or years, we will arrive at some consensus as to what is the best design.

I want to go back to some of the things that Vibeke covered. I think she covered the literature on therapies and radiographic analysis very thoroughly and very nicely. She pointed out the estimated progression rate or, if you will, an imputed progression rate based on historical data, which is a bit soft, was useful. But in my opinion, it's dangerous to try and use this to compare with the future course of a disease. I think you can use

this to compare treatment groups within the trial to show that they're more or less comparable, but to take this forward and say that treatment from this point on shows a difference from what we've imputed in the past is very dangerous and should not be done. I feel very strongly about that point. I just think it's not an appropriate method of analysis.

The other point I want to make, if I can take another minute or two, is that every trial enters a different population, number one.

Number two, everybody who reads films reads on a little different scale so that you can't compare readers in absolute terms. I think you can compare them generally in terms of progression over time.

The final point I want to make is we ought to be thinking rate, change over time, not absolute score. Absolute score will give you a little indication of what the severity of the disease is at the point of entry into the trial. But in terms of comparison of effectiveness of treatment, we need to be looking at rate of change. Now, that's usually accounted for by having a specific period of time, but I think we tend to be talking too much about absolute scores and we think, well, an absolute score of 40 in one trial is the same as an absolute score of 40 in another trial. That's not so.

DR. SIMON: Thank you, Dr. Sharp.

Bill?

DR. SCHWIETERMAN: Yes, thank you very much, Dr. Sharp. I thought that was helpful.

I just wanted to make one clarification or, at least, one point. The word "placebo" is often misused or, at least, used for many different purposes. I think since this is such a charged issue, we need to be talking about denial of standard of care or not because, in fact, placebos can be used on top of standard of care, as they were in ATTRACT study here. The agency position is that in fact equipoise has to exist in trials to begin with, and standard of care cannot be denied except in those circumstances where denial of that standard of care is inconsequential in the end. I don't want to get into all of that because I don't think we're here to discuss placebo-controlled trials.

DR. SHARP: I don't think we disagree.

DR. SCHWIETERMAN: Okay. I had been asked that question before. I just wanted to make that clear.

DR. SIEGEL: To expand a little bit on that, just to be clear, placebo can be used in active-controlled trials. You compare drug A to B and the people that get drug A also get a placebo for B. Those are not placebo-controlled trials.

This is a placebo-controlled trial. As Dr. Sharp pointed out, some people misunderstand that to mean that the control group was treated with placebo. All patients in this trial got methotrexate. The comparison was between getting the study active agent or a placebo. When we used the term placebo-controlled in the FDA regulations, that's what's meant. So, it's in that context we're calling it placebo-controlled, but we understand it's very different from not treating the patient.

DR. SHERRER: Is it really, though? Because in actual practice, if a person is on methotrexate and not responding, you're either going to change drug or add another drug, whereas you got them on placebo. So, in a sense, it is placebo because the standard of care, if you have suboptimal response to a drug, is to change that therapy in some way.

DR. SIEGEL: Well, that really addresses, at least in the construct that we think of things, the target population. The study by design is placebo-controlled because you're randomized to get placebo. The population are people who have had an inadequate response to methotrexate. One can then make your point to indicate is this an appropriate study or an appropriate population to study or an appropriate way to study them.

In general, except where it will mean harm to

the patient, we prefer, in studying a population that's failed to respond to a drug, that they be randomized to receive that drug again because experience has shown, like in, say, NSAIDs -- if you were to do a study where somebody didn't have a good response on NSAID and just compared another NSAID to placebo, I wouldn't be too surprised if any NSAID could be shown to work in nonresponders to other NSAIDs. But if you randomize them back to the one that they didn't have a good response to versus a new one, nobody has to my knowledge shown differences that way. So, there's a lot of inferential stuff about what you call a nonresponder.

DR. SIMON: Dr. Emery, then Barbara.

DR. EMERY: I perceive it as the wrong issue. The issue is active disease, and it's active disease over time. We designed the ATTRACT study several years ago. What I'm saying is ATTRACT is as good as it gets. You'll never be able to do ATTRACT again because we can't -- and I had a large number of patients in that study -- leave my patients with that degree of activity ethically anymore.

So, what we mustn't do is set hurdles which we're never going to achieve. We're not ever going to have the difference we saw between the two groups in any ethical study in the future. Therefore, the x-ray changes that we see between the two groups are going to be very, very

small.

Enbrel data, which we talked about this morning, two active groups. You see that effect.

If you have active therapy, x-ray is not sensitive enough to show differences that we can ethically allow to continue over a period of time that you're now talking about. So, if we're talking about what we're going to be setting for the next 3-5 years, we have to be realistic about what we can achieve between two active comparators because we can't have patients with active disease which is completely predictable now.

DR. SIEGEL: Well, of course, this study seems to have disproven that. You did show a difference between an active therapy and infliximab plus that active therapy. I don't know why one would presume that if a new class or a new really good agent came by, you might not show that the three-drug therapy is better than the two-drug therapy.

DR. EMERY: Because those patients now would go up to 25 milligrams of methotrexate intramuscular, would have two added drugs. There are many other biologic therapies. There are many other combination therapies that you can add. You don't leave patients with 50 milligrams of methotrexate. There are not those patients anymore.

I'm saying this for the sake of the development of new drugs and for biologics. You won't get new agents

otherwise.

DR. SIMON: Thank you, Dr. Emery.

Dr. White?

DR. WHITE: Just a couple comments because I've been thinking about all the trouble I had earlier when we were trying to think about "delay" versus "prevent" in the discussions. It seems to me that the discussions have assumed that a delay is worse than prevent. I actually want to raise that question because if delay is rate, which is what Dr. Sharp was talking about, that's really a measurement of rate, over what time, what changes do you get versus over what time, what changes do you get in another group.

It might be more important to have a drug that gave you a delay in all your patients than a drug that prevented no progression in 5 percent of your patients. One drug might get a prevent. One might get a delay, and clinically the delay medically would be a much more important issue.

So, from my way of thinking, just to lay it on the table, I think "delay" is a grand claim, particularly if it applies to most of the patients who have been treated. If you're going to go for "prevent," my own feeling is to me that would mean that you would want to have a definition of no progression that would have to take

in this error in measurement, and you would have to have no new erosions. To me that's no progression. And then you'd have to figure out how many people in each category fall into it and the reviewers would have to decide is that meaningful. You'll give us the statistics, but is that meaningful or not?

DR. SIMON: George?

DR. MILLS: In terms of all we've been hearing, what a wonderful situation, we're having a declining change, which is being assumed by everybody in the room now, and we're now saying we can't follow them for very long. I would like to draw back Dr. Sharp to the microphone and say exactly from the standpoint here of this time, this rate, how long would you like to be having us evaluate these patient studies knowing that, indeed, that variance, that difference is going to be smaller and smaller with the patient populations, as these drugs come available? It looks like to me we need relatively long studies to find a small change.

DR. SHARP: The statisticians can probably model this for you and give you a better answer than I can.

I think if you had a drug that produced almost uniform suppression of progression, instead of the amount of variation we have, we could find differences in smaller numbers of patients and perhaps in a shorter period of

time.

Now, I'm impressed with the data that Vibeke showed that there were a number of studies. We showed striking differences between an active drug and placebo or some comparator in 6 months. I must say that 20 years ago I thought it took 2 years to do a study. 10 years ago, I though it took 1 year. Now I'm convinced that you can do it in 6 months. We haven't pushed it back to 3 or 4 months. The MRI people are telling us that maybe we can, but they haven't proven it yet.

The duration of a trial I think really depends on the magnitude of change that you expect to induce and how consistent that change is going to be. Again, back to the point that Dr. Emery was making, if we're comparing treatments and we're looking at best treatment available today to compare the new agent against, our best combination of treatment to compare the new agent against, we've got to look at how consistent is that best treatment compared to what we're hoping to do. It's probably worthwhile to have a variety of treatments that are equivalent because some patients react one way or another to one drug and can't take drug A and can take drug B.

We don't yet understand exactly what failure to therapy really means. There are many ways of looking at what failure is. And we don't yet know whether switching

from drug A to drug B really is the appropriate thing to do when we have "failure". We haven't really defined failure. It may well be, but it's worthwhile switching patients that have failed, but we haven't proven that yet.

I'm sort of wandering I'm afraid. Have I answered your question?

DR. MILLS: You've answered the question the way I expected it. From the standpoint here, it's to draw out from you that, indeed, the time interval may shorten because we may be looking at more soft tissue changes, more cartilage changes, and you're alluding to potentially having to use a different monitoring device, such as MRI, versus standard posterior/anterior radiographs for sensitivity to pick up these changes. So, we need another modality, it sounds like, to be able to assist us in looking at these trial intervals over duration, especially when there's a very small amount of change you're suggesting.

DR. SHARP: Well, personally I think that in the next decade at some point, we're going to be looking at synovitis in terms of a predictor of what happens structurally. Again, we haven't proven that the technology can be used in an organized fashion to prove that.

I think in a schematic, one ought to think of the process of rheumatoid arthritis, which is inflammation,

producing the outcome, which is damaged bone and cartilage and ligaments and tendons and so forth and that, all together, producing disability. Now, if we can stop the inflammation and we have a way of proving it, if we really know what's going on in the histopathology of the disease -- and I think we do -- and if we can measure that accurately, then we ought to be able to measure synovitis, show that it goes away, and know that we're preventing disability and deformities and structural damage.

DR. SIMON: David?

DR. WOFSY: I'd like to take a shot at viewing this from the public perspective. The reason for having different designations for a guidance document laying out different possibilities that one could delay or one could prevent is because one would be preferable to the other and you'd be giving people useful information to distinguish between drugs.

My own view is that when this document was in composition and there were no drugs that had demonstrated elay or prevention, there was sort of a basis for thinking ut that kind of a distinction. But now here we are. We a document that theoretically could lead to drug A aid to prevent, drug B being said to delay, people stinctions between them on that basis, even though 's olutely no compelling evidence in any way to say

that they would be different. That would mislead the public.

So, it seems to me, now that we really are dealing with agents that have this effect, we have to revisit that point. Do these guidance criteria really imply a meaningful difference to the public, communicating information that's real? It seems to me that what we are now dealing with in real life is that they don't.

The only way to do that, for some of the reasons that have just been described, the sort of small differences, the difficulty doing placebo-controlled trials, is if somebody thinks they have a drug that in fact is better than someone else's drug in the degree to which it will affect structural damage, there needs to be a head-to-head comparison. I think what we've heard, all of this discussion of the last 20 minutes says that without a head-to-head comparison, we will never be able to distinguish between these agents that have now been shown to have an effect on structural damage. We'll never be able to distinguish between them. If we get sucked into using language that distinguishes between them, the language won't reflect the reality.

DR. FIRESTEIN: Yes. I just wanted to come back to one point that Dr. Sharp made about understanding the pathogenesis, the disease, and us looking earlier and

earlier, and then somehow measuring synovitis. Again, I remind people that the pathogenesis of bone and cartilage destruction is extremely complex, and I think it is well beyond our current technology, in imaging especially, to be able to say that the synovium shrinks or that we somehow can image synovitis and that will be useful as a surrogate marker, for instance, for later radiographic damage.

There are many ways that we can make joints smaller, with anti-inflammatories, for instance, that have no effect on the progression of structural damage. Some of the mechanisms in very late disease for joint destruction can be different from the mechanisms in early disease in terms of structural instability. They can contribute to it and other mechanisms of joint destruction.

So, I think we have to be careful we don't go along a pathway where we believe that we can learn as much as we need to know in 3 or 4 months by doing MRIs to look at the bulk of the synovium, for instance, how big that is and how much we shrink it down.

DR. SIMON: In extending that, it's very important to remember that years ago there was actually a large study done. There was a group of patients with rheumatoid arthritis that by pathology don't have very much synovitis but have equally just as much destruction, and that was considered a fibrotic form of the inflammatory

process. So, measuring by volume synovial pannus in that particular patient population would be misleading because, in fact, their destructive potential was just as great.

Is that the same disease? Well, we don't know. We use the same criteria to define that disease, but we may now actually be bordering on the final, real recognition that what we call rheumatoid arthritis is a group of patients with a heterogeneous process based on genes, gene response, and any number of other things.

I presume you have one more comment to make?

DR. SHARP: One more comment, yes.

(Laughter.)

DR. SHARP: In response to Dr. Firestein's comments. Nirvana isn't here yet. I agree.

You can take my remarks as where I think we should be going and doing appropriate studies with the expectation that we probably can develop methods. Now, will we? I'm predicting we will, but it has to be done.

DR. WINALSKI: I would agree with both of you in that I think there needs to be a disconnect between synovitis and bone changes. I think that's clear from the ACR20 response compared to the radiographic progression, just to start. MR can measure these things right now, but we don't have as tight errors as we want or we need. If it's taken 20 or 30 years to get this far with the Sharp

and Larsen scores, I think it's going to take a long time with MR as well.

Also, in that vein, I think that those are the sorts of studies we need to be doing because I think we're in the unenviable position that we now have drugs that are making us need these tight error bars on our measurements. I think we're at a point now where, to differentiate between two excellent drugs, you either have to have a long-term study, which is corporately undesirable, or you need to get a better measurement.

DR. FIRESTEIN: The one thing I would add to that is there's been a lot of discussion about how tight the error bars have gotten and how much better things are, but I would just remind people that the response rates, even for these outstanding new agents is on the order of 60 to 70 percent and only about half of those meet ACR50 criteria. So, there is still a huge unmet medical need in terms of our rheumatoid arthritis patients. I think we are clearly far ahead of where we were 5 and 10 years ago, but there is still a lot of room for improvement.

One other small point I wanted to make -- and this was brought up earlier and I forgot who had mentioned it, but this notion of being able to look at one's progression of disease radiographically and then being able to predict where they would go if they weren't treated as

sort of a rough way of doing a self-controlled study. It's clear from a number of studies that have been published, looking at individual patients over time, that there is a broad variety in terms of the courses that they follow, anywhere from linear to flat progression, followed by acceleration to early erosions, followed by flat disease over a period of up to 10 years later. One has to be very careful about using that sort of retrospective analysis in a radiographic study.

DR. SIMON: Which then brings us really to the specific questions. I think we really dealt with most of the things in the first part, but one of the things that we've not yet dealt with -- I personally have some very strong opinions about this, but I won't state them yet -- surprise -- is, are there criteria available to select patient populations who are likely without treatment to develop erosions? What do you think?

DR. FIRESTEIN: Yes.

DR. SIMON: Do you want to elaborate on that?

DR. FIRESTEIN: Well, I don't know what they all are. I'm hoping sometime in the next 5 to 10 years with genomics and microarray that we'll know a lot more.

But certainly one can predict that seropositive patients, patients that have probably the susceptibility cassette, have nodules, are more likely to go on to have

erosive disease.

DR. SIMON: So, the assumption there is that it is different diseases and that there are patients who are more likely to have erosions, and thus they have ishkabibble rheumatoid arthritis, and there's another group of people that aren't likely to get erosions. They have a different genomic background most likely, and they don't have ishkabibble rheumatoid arthritis.

DR. FIRESTEIN: I probably wouldn't term it exactly that way.

(Laughter.)

DR. FIRESTEIN: It clearly is a heterogeneous disease, but within that large number of diseases that we call rheumatoid arthritis, there are some broader characteristics and specifically seropositive patients comprise about 85 percent. If you look within certain, distinct ethnic and racial groups, you can find a very high percentage of patients that have the susceptibility cassette in them. That really starts to lend credence to the notion that within this morass there is some homogeneity that you can pull out that maybe comprises the majority of patients, and then there's all the people around the fringe.

DR. SIMON: So, I'd like to move on to the second question which turns the table entirely. We've been

talking about no new erosions. Now the question is could you envision a circumstance, and thus envision what you'd call it, where you actually reverse erosions that are present?

A rose by any other name, that suggests healing. Healing of disease then has two components to it. One is erosions which are stated here, and the other one is reconstruction of whatever it is that's associated with losing joint space, thus recreating articular cartilage, which is what makes up your joint space.

So, how do people feel about the idea of is a reduction in number of existing erosions a viable claim in rheumatoid arthritis? Dr. Katona?

DR. KATONA: I would like to take this opportunity to combine question number 1 and question number 2. One of the previous comments during the discussion previously which was the most appealing to me was that if we look at the field from the clinical point of view, the best hope for us would be that there would be a wide array of drugs which would not be that much differentiated by the label since I really don't think 3 months, 6 months, or a year could differentiate what a drug does in a 20- or 30-year disease.

So, if we say that we would set a short-term goal for us to let the companies do the best work and have

a relatively comparable labeling at the beginning, but then ask them to require to do long-term studies and this is what would come out of it, is this possible?

I think one day, hopefully, a drug will come around where it will be possible. There is no way that we're going to know anything in 6 months or in a year. And anybody wants to develop a claim -- I don't know whether it's wise from the agency point of view to tell them not to try because maybe there will be a drug which is going to be a wonderful drug and going to answer all our clinical problems. But that would be my long-term view of this whole process.

DR. SIMON: Dr. White?

DR. WHITE: I was really taken aback when we saw Dr. Maini's slide and his animal in which he showed us just that and showed us really stunning histologic pictures. To me what I really wanted to hear the science of was what was the science of that recovery and taking the joint that was inflamed to a joint that now had cartilage on it. That's what he showed us.

So, I think that if I were a patient, that's the one I would want. I would want the healing one. That's the real bottom line.

So, I would like to be able to have that claim be possible. I would sure like to encourage companies to

go after that. From a patient standpoint, I couldn't imagine what more you would want than to take a joint that was damaged and to reverse that. I don't think it's so far fetched seeing the animal model, the data that we saw.

DR. SIMON: Gary?

DR. FIRESTEIN: With regard to that model, it's a very elegant model. One of the things you have to be careful about is the age of the animals in that particular study. I don't remember the exact age. They were, I think, 3 weeks old when treatment was initiated, in that general vicinity. The cartilage and mesenchymal tissue is much more plastic during that period of time compared with adult animals or even adult humans. So, I think there's no question that one can develop new cortical bone that can be involved with healing erosions in adults, but I think one has to be very careful about extrapolating from neonatal animal data to growing new cartilage, for instance, in adult patients.

DR. SIMON: Yvonne?

DR. SHERRER: I was just going to make the point that I think from a patient's point of view and a clinician's point of view, you wouldn't want that erosion data in a vacuum. You would want the company to show that along with the "healing" of erosions, you had sustained improvement in disability over time, sustained improvement

in pain over time, and that patients got back to excellent quality of life over time. I think that's what patients want to see given that you can't do to them what you could do to the mice.

DR. SIMON: As Dr. Sharp said as he was leaving, this is an issue of Nirvana. On the other hand, the FDA was actually quite practical when they wrote this question, and the next part of this question is really related to how in the hell are you going to prove that and what does it mean. So, obviously, we're not going to be able to do biopsies on people to demonstrate regrowth of subchondral bone, bone and cortical bone and then cartilage.

So, the question that they want us to take an extension on this -- yes, obviously we want to have drugs and therapeutic interventions that cure the disease, that put us back the way we were. How are we going to prove that?

You're asking please discuss the ways in which these outcomes could be measured, which imaging modalities, duration of study to determine durability of effect might be there. And is there a minimum number of erosions -- that's a very interesting idea -- compared to baseline that should be healed in order to consider a product reasonably likely -- reasonably likely -- to confer clinical benefit?

Almost talmudic in nature.

(Laughter.)

DR. SIMON: So, the real question here is would an MRI study do that? And do you think that MRI could eventually do that?

DR. WINALSKI: I think that the definition of an erosion is going to become critical because right now when I think of just radiographs now, is healing of an erosion restoration of a pristine subchondral bone or is it cortication of a previous erosion? What's the difference between a subchondral cyst and an erosion? That's in radiographs.

When you get to MR, you now have what we roughly call bone marrow edema, which is an edema-like signal. Whether it's truly edema of the marrow or whether it's increased vascularity or exactly what it represents, we don't know. And I can imagine restoration of that marrow signal to be one thing to look for, but if you actually have loss of the cortical bone and the subchondral bone, that's I think a more difficult call on MR than it is on radiographs even.

So, I think that it's going to be difficult, but I think there can be definitions set up so that everybody is talking on the same page, which is the important thing.

DR. SIMON: Mark?

DR. SCHWEITZER: Yes, I would second call. I really have no doubt that once an adequate baseline database for various types of arthritis, particularly rheumatoid, is in the literature that MR will be the way to evaluate these patients. I think we've gotten a long way there, but we still have a way to go before it's an adequate database.

But again, I'm going to be a splitter. I'm going to say on MR you're going to look at a whole bunch of different things. You're going to see erosion/geodes in much larger numbers than we're used to seeing radiographically. So, therefore, the changes are probably going to be easier to discern than radiographically because you're dealing with bigger numbers. You've got to look, as Carl said, at the marrow edema type signal. With small joints in the hand, it will be still be hard to look at cartilage, but I think in maybe two years we'll be there. We're going to look at joint fluid volumes. You're going to look at tendons, something that we haven't looked at yet, and look at synovial volumes.

Again, I will agree that in all situations the volume of synovium itself is not the be all and end all of a marker for disease, but just splitting all of these different things to look at, it is certainly one thing to

look at. That probably will be the only modality which we'll be able to split all these things and see what disease-modifying agents modify what aspects of the disease.

DR. SIMON: So, the ultrasonographers of the joint will take issue with that.

Dr. Emery?

DR. EMERY: I just agree with what has been said, but the definition we've used for erosions for MRI -- and we've now done a five-center validation of this -- with the weighted cappers for erosion is about .96. Very high indeed for the best films and MCPs. So, we're pretty good actually at distinguishing erosions, which actually are defined on two planes involving the cortex, clear margins, and there are clear definitions of those.

Around them are these T-2 fat-suppressed edematous lesions. The problem with healing is the first thing you do with any of these therapies, methotrexate or the biologics, when you give it to them is you lose the edema. Then the hole takes much, much longer. The hole we know on ultrasound and on x-ray is what an erosion is. We've biopsied it and we've got year data on this now. But to heal that erosion takes a very long time because you're going from a very active process, because most of these are untreated when they go in, to actually getting an

osteoblast to form over the top of this. To date, it's very difficult to look at.

But what is absolutely clear, if you want to know if an erosion heals looking at comparative MR, x-ray, ultrasound, you've got to have a three-dimensional technique. Because of the variability of an individual erosion, you can't tell.

The problems exist when you get big lesions which then contract into two separate ones, how do you define that when an erosion seems to get two separate areas in it? But these are being worked out, and I think the international collaborations in MR now are actually taking things a long way forward. As I say, we've now got longitudinal data which are actually scoring these which I think will make a major advance.

DR. WOFSY: I think it's a good sign for our field that we're beginning to talk about some of these technologic advances, but I don't think this discussion is unique to us. It's happened in every other field.

The important principle that's been established in the last 3 years is that there are agents that can affect the progression of structural damage in these diseases. That wasn't known 3 years ago. Now there are five drugs for which it has been proven. Maybe more. I can count five quickly.

Now, the technology is going to change. It's going to change every year just like it does with cardiac cath and everything else. So, the gold standard by which we judge whether something improves the joint is going to be different every year. We've already proven that we can't anticipate it and there's no reason to think we can anticipate it today for 3 years from now any better. We'll always have a different technology, and the FDA should always expect state-of-the-art technology. What we got today was state-of-the-art technology. I didn't hear anybody fault it and I certainly wouldn't fault it. But it becomes trees instead of forests at a certain point. Next year it may be MRI.

The important point is the establishment of the fundamental impact on the biology of the disease, which we didn't know about before. Now I think the reason we're quibbling over some of this is the hope that the FDA, by choice of sort of what the latest technology is, the latest study that came along, will appear to make distinctions among these agents.

I'm hitting the same note that I hit before. I'm going to try it one more time. I think that's an unreasonable thing for the FDA to do. I think it's an unreasonable thing for people to ask the FDA to do. At a certain point, if you think your approved drug is better

than their approved drug, you invest in a comparison. We have now move beyond the first question where you could do placebo-controlled trials to the second question where the head-to-head comparisons are what people need to know. You can't be turning to some other methodology to try to make those distinctions.

My own view here is now you have a lot of agents that are proven their value. You will have more agents that prove their value. You want to compare them? Compare them.

DR. SIEGEL: I do want to say one thing about that. I certainly agree and couldn't agree more that the only valid way to study comparisons is with a direct comparison. We certainly wouldn't think of approving any comparative claims either in the labeling or in marketing that were not appropriately based on comparative studies.

But what happens in many fields in medicine -- it may or may not have happened here -- is notwithstanding the fact that you can't support a difference, labeling can differ, and labeling can differ because people do different studies. They study different stages of the disease, different severities of the disease, different combinations, and different endpoints. So, the fact that you can't make a comparative claim and shouldn't doesn't mean that the labeling can or should be identical because

the labeling does have to reflect what has or has not been shown individually for a given drug.

DR. SIMON: I think that Dr. Wofsy was really inferring the idea that certain words are chosen based on being very dynamite loaded and can be then used to distinguish among products or between products. Sometimes those are just that: words. And there really is no difference.

DR. SIEGEL: Right. And we are sensitive. We've seen companies go out and say we're the only company who's approved for doing this.

DR. WOFSY: Actually I think these comments -- and I want to be very explicit about this -- are appropriate for this afternoon's discussion, although this discussion absolutely takes place sort of in the context of the specific issue we were dealing with this morning and knowing that it's still hanging unresolved.

I don't speak up to speak to what should or should not be done in this case where there's a history and there are agreements between the FDA and the sponsor and a whole set of things that I think influence this issue. It has more to do with the future and to whether or not the way to deal with this problem is to make the next generation guidance document with its language.

Really what I'm just saying is that there's I

think a lesson in the dilemma we're facing today, and the lesson in part is that the next time the language is rewritten in anticipation of whether it will be MRI or who knows what technology to do it, we will probably find ourselves 3 years down the line in the same dilemma.

DR. SIEGEL: Which doesn't, of course, mean that it shouldn't be done. As someone who only contributed an extremely small amount to the development of this guidance, I want to say that while we can look in retrospect and say knowing what we know now, there are things that would be written differently, I think the data would show, without drawing a clear-cut causal relationship, that the guidance at least allowed the carrying out of several extremely well-designed, useful trials in this area.

That's one of the things about giving guidance. You'd like to know all the answers before you give guidance, but if wait until you know all the answers, it doesn't get you anywhere. So, we're in this cyclic thing.

DR. SIMON: Dr. Wolfe?

DR. WOLFE: I want to say a word in behalf of longitudinal observational studies because the world has shown that the results of randomized, controlled trials don't always work out as well in real life. There's the wonderful observation about auranofin of a number of years

ago, which passed with flying colors through everything and turned out to be a drug of not much use.

In addition, patients in real life don't do as well as they do in randomized trials. There are more dropouts. It's been shown over and over again.

I think what really needs to be done is to take the information that one gathers from trials like this and insist that people do long-term follow-up studies with radiographs and outcomes because that's the only way you're really going to know whether these drugs work for more than the period of observation and work for more than in just a very selected population.

I want to make another point about radiographs for a moment, which is that radiographic progression is a function of disease activity. It has been shown repeatedly that the higher the C-reactive protein, the more likely you are to have erosions.

Now, when you come in to trials like this, people have very active disease, and they are not representative of most rheumatoid arthritis patients. In general, as I've looked across these trials, the overall activity on a percentile scale is close to 70 percent of maximum disease activity for patients with rheumatoid arthritis. You can't get ACR50's and very many ACR20's once you begin to reduce your disease activity much below

the median. You have to have really active disease to show these changes. But most of our patients, most of the people we treat are very different from the patients that have been seen in all of these trials.

The second thing I think that we ought to attempt to address sometime is treating the average rheumatoid arthritis patient. The average rheumatoid arthritis patient also becomes disabled and dies early. One of the things we need to do is see whether we can apply the methods that we're applying to very severe patients to those who have less severe disease and represent the majority of rheumatoid arthritis patients.

DR. SIMON: Dr. Katona.

DR. KATONA: I have had the privilege of participating in the discussions now for about 2 years. Going back to the very young mice, the TNF-alpha congenic mice, who improved so much, I am just delighted that most of the companies present a lot of pediatric data on these discussions. I was somewhat disappointed today that I did not hear any pediatric data presented on this particular drug.

But I just wanted to talk with the agency about what is the status of the pediatric studies. I know that most of the time there is at least PK data, but future plans and follow-ups as well as special emphasis. Because,

as you know, we have just as much problems with the kids and the effects on children will last for not only 20-30 years, but 70 and 80 years. I really, really would like to encourage both the agency, as well as the sponsors, to take very seriously the pediatric population.

DR. WEISS: Just a real quick comment. Pediatric initiatives have been extremely important at the agency in actually all three centers, but in the Center for Biologics and the Center for Drugs, a lot of efforts are going on to encourage studies certainly in diseases that also affect pediatric patients. You heard from the sponsor that they are pursuing pediatric data in not only Crohn's disease, which was their first approved indication for infliximab, but also in JRA.

As is the case right now, many of those studies tend to lag behind the adult data for a number of different reasons, sometimes just the fact that the numbers are different, sometimes the fact that it's important to know, first of all, whether it works sometimes in adults, the proof of efficacy, before going on and actually studying pediatric patients or a more vulnerable population. So, for a number of different reasons, there are oftentimes delays.

For some very serious diseases, there oftentimes are not delays and we encourage getting

pediatric data in HIV and other settings sometimes almost simultaneously with the adult data or certainly after the phase I data, but in other diseases there are delays. But those studies are ongoing or in very active discussions with the agency.

DR. SIMON: So, I think that we have gone through the process of looking at the questions that you asked. Was there anything else that came up during the discussion that you'd like to now address, Bill or Jay?

DR. SCHWIETERMAN: There's plenty that came up during the discussions that haven't been addressed. I don't have any illusions about this being easy, but I'd just like to thank you and the committee for what I think has been a very helpful introduction. We plan on taking this under advisement and will certainly keep this committee apprised as to how we go from here.

DR. SIMON: I'd like to thank the committee.

Does anybody else have any other comments to make?

(No response.)

DR. SIMON: Thank you very much for a wonderful meeting.

We stand adjourned.

(Whereupon, at 3:55 p.m., the committee was adjourned.)