FOOD AND DRUG ADMINISTRATION

CENTER FOR DRUG EVALUATION AND RESEARCH

+ + + + +

ARTHRITIS ADVISORY COMMITTEE

+ + + + +

WEDNESDAY,

MARCH 5, 2003

+ + + + +

The above-entitled meeting was convened in the Kennedy Grand Ballroom of the Holiday Inn Silver Spring, 8777 Georgia Avenue, Silver Spring, Maryland, at 8:00 a.m., Dr. Steven B. Abramson, Acting Chairperson, presiding.

PRESENT:

STEVEN B. ABRAMSON, M.D., Acting Chairperson

KATHLEEN REEDY, R.D.H., M.S., Executive Secretary

JENNIFER ANDERSON, Ph.D., Member

SUSAN M. MANZI, M.D., Member

H. JAMES WILLIAMS, Jr., M.D., Member

WENDY W. McBRAIR, R.N., M.S., C.H.E.S., Consumer Representative

ARTHRITIS ADVISORY COMMITTEE FDA CONSULTANTS:

STEVEN B. ABRAMSON, M.D.

KENNETH D. BRANDT, M.D.

JANET D. ELASHOFF, Ph.D.

JAMES F. FRIES, M.D.

ALLAN GIBOFSKY, M.D., J.D.

ROBERT W. MAKUCH, Ph.D.

FDA CONSULTANTS FROM OTHER ADVISORY COMMITTEES:

RUTH S. DAY, Ph.D.

JAMES H. LEWIS, M.D.

LEONARD B. SEEFF, M.D.

C O N T E N T S

C O N T E N T S (Continued)

PAGE

P R O C E E D I N G S

(8:07 a.m.)

CHAIRPERSON ABRAMSON: We'd like to call the meeting to order, please.

I am Dr. Abramson of NYU and the Hospital for Joint Diseases.

And we'll begin the meeting by having the committee introduce themselves, and we'll begin with Dr. Seeff, please.

DR. SEEFF: Leonard Seeff from the National Institutes of Diabetes and Digestive and Kidney Diseases, NIH.

DR. LEWIS: I'm James Lewis, hepatologist at Georgetown University.

DR. DAY: I'm Ruth Day from Duke University and a member of the Direct Safety and Risk Management Advisory Committee.

DR. FRIES: Jim Fries, Stanford University rheumatologist.

DR. BRANDT: Ken Brandt, rheumatologist, Indiana University.

DR. ELASHOFF: Janet Elashoff,

biostatistics, UCLA and Cedar Sinai.

DR. MAKUCH: Robert Makuch, head of biostatistics, Yale University School of Medicine.

DR. ANDERSON: Jennifer Anderson, statistician, Boston University School of Medicine.

MS. McBRAIR: Wendy McBrair, Director of Arthritis Services, Virtual Health of New Jersey, consumer rep.

DR. WILLIAMS: James Williams, rheumatologist, University of Utah.

MS. REEDY: Kathleen Reedy, Advisory Committees, Food and Drug Administration.

DR. GIBOFSKY: Allan Gibofsky, rheumatologist, Hospital for Special Surgery at Cornell in New York.

DR. GOLDKIND: Larry Goldkind, Deputy Division Director at Division of Anti-inflammatory, Analgesic and Ophthalmologic Drug Products.

DR. SIMON: Lee Simon, Division Director of Analgesic, Anti-inflammatory and Ophthalmologic Drug Products and a rheumatologist.

DR. BULL: Jonca Bull, Director, Office

of Drug Evaluation V in the Office of New Drugs.

DR. KWEDER: I'm Sandra Kweder, the Deputy Director of the Office of New Drugs.

DR. WOODCOCK: Janet Woodcock, head of the Center for Drugs.

CHAIRPERSON ABRAMSON: Thank you.

We'll now have a meeting statement read by Ms. Kathleen Reedy, Executive Secretary.

MS. REEDY: For the Arthritis Drugs Advisory Committee on March 5th, 2003, addressing Arava, leflunomide.

The following announcement addresses the issue of conflict of interest with regard to this meeting and is made a part of the record to preclude even the appearance of such at this meeting. Based on the submitted agenda for the meeting and all financial interests reported by the committee participants, it has been determined that all interests in firms regulated by the Center for Drug Evaluation and Research present no potential for an appearance of a conflict of interest at this meeting with the following exceptions.

Full waivers have been granted to the following participants in accordance with 18 United States Code 208(b)(3):

Dr. James Lewis for serving on a competitor's speakers bureau. He receives less than $10,001 per year and lectures on topics unrelated to Arava or its competing products. The waiver also includes his consulting for the sponsor on issues unrelated to Arava. He receives less than $10,001 per year.

Dr. Kenneth Brandt for consulting for the sponsor on unrelated issues. He receives less than $10,001 per year. For consulting and lecturing for a competitor on unrelated issues, he receives between 10,001 and $50,000 per year.

In accordance with 18 United States Code 208(b)(3) and 505(n)(4), Dr. Allan Gibofsky for ownership of stock in two competitors, one stock valued between 5,000 and 25,000 and the other valued between 25,001 and 50,000.

For consulting for three competitors for which he receives less than $10,001 per firm per

year and for lecturing for three competitors for which he receives less than $10,001 per firm per year, Dr. Gibofsky's consulting and lecturing is unrelated to the competing products.

A copy of the waiver statements may be obtained by submitting a written request to the agency's Freedom of Information Office, Room 12A30 of the Parklawn Building.

Dr. John Cush has been excluded from participating in today's discussions due to his current involvement in studies on two of the competing products and his past consulting on the product at issue.

In the event that the discussions involve any other products or firms not already on the agenda for which an FDA participant has a financial interest, the participants are aware of the need to exclude themselves from such 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 products they may wish to comment upon.

CHAIRPERSON ABRAMSON: Thank you.

Today's meeting will be on the recent update on the efficacy and safety of Arava or leflunomide, and the first presentation will be by Dr. Simon on the regulatory history, Arava and treatment of rheumatoid arthritis.

Dr. Simon.

DR. SIMON: Is that actually now how the agenda is supposed to go?

CHAIRPERSON ABRAMSON: I apologize.

DR. SIMON: Excuse me. That's okay.

So basically I'm here to welcome you all first and to go over the agenda briefly, and I would like to welcome you all in the name of the agency, and thank you from the bottom of my heart for the division, that you had to reach 700 pages of briefing documentation prior to coming here. We are quite grateful that you've taken the time out of your busy schedule to be able to offer us your

advice on this particular thorny issue, and I will review with you what we're going to be doing today through a review of the agenda. Although you have a printed agenda in front of you, basically this is a little bit more detailed.

I will in a few minutes begin with a regulatory history of Arava in the context of therapy for rheumatoid arthritis, and then we're going to move on to a discussion of outcome measures for disability and physical function.

There will then be a sponsor presentation of efficacy.

There will also be an FDA statistician's assessment of impact of placebo withdrawals in the two year landmark analyses for improvement in physical function, and this will be representing the meat of the data for a discussion regarding change in the guidance related to two to five years of efficacy data for the indication for improvement in physical function.

A discussion of questions regarding efficacy of Arava in the context of the indication

for the improvement in physical function.

Then a discussion of the RA guidance document of 1999 and the indication for improvement in disability which presently requires the two to five years of data.

In the context of this afternoon, we're going to have an FDA presentation regarding hepatotoxicity associated with Arava; the sponsor presentation of overall safety of Arava and its benefit to risk ratio for use in the context of the universe of therapies for rheumatoid arthritis; a presentation regarding risk communication, how one conveys potential risk, which I think you'll find very interesting, and then the further discussion of questions.

As noted, not on the agenda, but on your printed agenda, we have two periods for open public comment, each of which one will be in the morning and one will be in the afternoon.

Thank you, Mr. Chairman.

CHAIRPERSON ABRAMSON: Thank you. Thank you, Lee.

So we will go to the open public hearing, and we'll have a statement first by Dr. Sidney Wolfe.

DR. SIDNEY WOLFE: Thank you.

I'm just going to talk for a few minutes now, and most of my comments will be in the public hearing in the afternoon on the safety issue.

I used both in our original petition to take leflunomide off the market and in preparing comments for today the FDA medical officer's reports, which give slightly different results in terms of effectiveness. In the M302 study, as you know, the largest of the studies with roughly 500 people in each leg randomized to get methotrexate or leflunomide, methotrexate was significantly better.

In the other two there was really not a significant difference between them, and in MN301, as you know, leflunomide and sulfasalazine were roughly about the same.

So the statement that we made, which Dr. Simon seems to rebut in his comments, that there was no evidence that leflunomide was any -- offers no

advantage to patients with rheumatoid arthritis compared with methotrexate, which is obviously the context, the statement is correct, and I don't know why it's labeled as inaccurate evidence.

I will mention now and in more detail this afternoon the fact that this is the first time in, I guess, the 32 years that I've been monitoring with my group, Public Citizen's Health Research Group, the FDA and the pharmaceutical industry that I've ever been asked to do something by an FDA Advisory Committee member.

Dr. David Yocum is the one that said he had had a tragic death from hepatic necrosis in a patient using this drug. It had a hypertensive episode and a stroke in another patient and personally stopped using the drug and literally called me and asked me if we would consider a petition to take it off the market.

The more I learned about it after his call, the more I was convinced. I did talk with him a couple of days ago to see whether he has still stuck by his guns, and he says he still does not use

this drug. He finds it's entirely possible to practice good rheumatology without this drug.

Like others, he starts with methotrexate first, which is as effective as I will discuss this afternoon, safer and certainly less expensive than either the TNF modifying drugs or leflunomide.

I would just also like to comment for a minute on this idea put forth by Dr. Simon that it isn't possible to do two year randomized controlled trials to look at disability. We certainly in a number of other spheres with people who are probably more mobile literally in terms of moving around or whatever than a lot of people with rheumatoid arthritis have been able to do two or longer year trials in hypertension, the Women's Health Initiative trial, and so I don't understand why it isn't possible to do it here.

And in some of the data that Dr. Simon is presenting, the patient accountability section, it looks as though to me that at the end of two years more -- I don't know if it's quite statistically significant -- but certainly more

people completed the two years on methotrexate than did on leflunomide.

I don't see why one has to -- I mean, I understand the attractiveness and the simplicity of the scales that Dr. Fries has worked on for a long time, but I think that they really are not a substitute for good epidemiologically derived data from randomized controlled trials, and I think it's possible to do that, and I think that should continue to be the goal to go for, not to try and make distinctions that I think are without a difference based on a scale that is really of lesser validity.

This is really all I have to say this morning, and again, this afternoon I will present a much longer amount of information on the hepatotoxicity and other kinds of toxicity, and I again thank you for the chance to speak for a few minutes this morning.

CHAIRPERSON ABRAMSON: Thank you very much.

DR. SIDNEY WOLFE: Do you have any

questions for me?

CHAIRPERSON ABRAMSON: No, there will be no questions.

The next speaker is Mr. Kevin Brennan, Senior Vice President, Health Policy of the Arthritis Foundation.

While we are waiting to see if Mr. Brennan is here, there are two statements that Kathleen Reedy has received that would like to be entered into this open segment.

MS. REEDY: This is from Ray Timmons.

"I hear that FDA is meeting to discuss Arava. It has been a miracle for me. Without, I would probably be in a wheel chair and out of work. Please keep it on the market!

"All DMARDs have a risk of death. If you look at the studies carefully, Arava has no more risks than others. The only study that seemed to indicate otherwise (the one in Europe) showed other factors (such as already damaged liver or other liver damaging drug) in all except one death. It only showed that Arava taken with something else

that also damages the liver is dangerous.

"The practice I went to a few years ago had two people die from take methotrexate in one month. If you did a study comparing methotrexate to Arava, you will find Arava to be much safer and much more effective. It just that methotrexate deaths are no longer being reported."

And this is a patient obviously.

And another patient: "I understand Arava's benefits are under question.

"I was on the drug study for Leflunomide that was later marketed as Arava. I have Rheumatoid Arthritis sine 1980 & have been through many medications. I have Tinitis caused by the Aspirin in so many of the meds. Esophagitis & other stomach problems because of side effects of some of the meds. I was on one for 12-1/2 years & woke up one morning realizing it no longer worked.

"Arava not only helped the inflammation (sic) & pain, it was kinder to my stomach & produced no other side effects. I am now on 20 mg. daily & doing very well. I'm sure this is a medication

doing much more good than harm. Hopefully it will continue to help me. I'm now 77 years young."

This woman's name is Dorothy Karo.

CHAIRPERSON ABRAMSON: Okay. Thank you. Is Mr. Brennan here?

(No response.)

CHAIRPERSON ABRAMSON: All right. If not, we'll go back to the agenda and reintroduce Dr. Simon on the regulatory history of Arava.

DR. SIMON: We always like surprises. So good morning again, and I want to thank the commentator for the open public forum this morning so far, and he has raised several issues that I will address in my presentation.

We thought it would be cogent to sit down and recognize where we are in the treatment of rheumatoid arthritis today and where we came from. Again, as I mentioned in my introduction, I am a rheumatologist. I was in practice for over 20 years in Boston, and I continue to speak to my patients periodically even though I'm here now at the FDA.

In general, my role at the FDA has

allowed me the privilege of looking at the treatment of the diseases that I was occupied with as a clinician in a very different way than before, and I hope that my presentation today will somehow reflect that for you.

So in thinking about what is rheumatoid arthritis, and I hope that the committee will bear with me because, in fact, there may be some people that were here today that were not here yesterday, and there may be some people who aren't as evidenced about rheumatoid arthritis as others around the committee.

So rheumatoid arthritis is a disease that affects about one percent of the U.S. patient population, and although it can affect anyone at any age, the peak onset is between the ages of 20 and 50, which is the most productive years of one's life, although now that people are living well past 100, it's not to suggest that people can't be productive after that as well.

It is a heterogeneous disease with a clear variable course. It's a systemic inflammatory

disease associated with an as yet poorly understood immune dysfunction; leads to the development of destructive erosive disease in a great majority and remissions are rare. Cure has not yet been observed.

As you heard, yesterday it shortens life span in some patients. The clinical outcomes are most notable for the state of debility. So the idea is to prevent debility. The idea is to be aggressive in treating the systemic inflammatory disease to prevent these events from taking place.

Other questions have arisen about certain other issues some of which you've heard about yesterday as well. So there are questions regarding disease and an increase in cardiovascular events associated with this disease. There's also questions about the incidence associated with rheumatoid arthritis, both treated and untreated for non-Hodgkin's lymphoma and other forms of malignancies.

But, in general, as per the last bullet, most patients suffer an unrelenting course. It's

characterized by recurrent flares over years leading to progressive loss of functional status and ultimately leading to significant disability. An unfortunate few have an accelerated mutilating course and another lucky few have either mild disease or enter into remission early.

So this chronic inflammatory autoimmune disease begins in the synovial membrane and then subsequently over time not only affects the cartilage and bone and soft tissues of the joint, but also affects extra-articular sites establishing a systemic disease as well as a joint disease. It is in some fashion associated with the presence of rheumatoid factor, which is an autoantibody, and that may be epiphenomenal or actually may be causal in some people's lexicon.

It has a clear genetic predisposition with a familial incidence. We now know that HLA-DR4 related antigens are clearly associated with the onset of worse disease, and there is as yet an unknown environmental trigger perhaps a virus, a ubiquitous disease that affects the specific genetic

host. We yet don't know.

This cartoon demonstrates the complexity of the events that take place in leading to the destructive lesion that we know of here at the joint level. It begins with an antigen presenting cell interacting with a T cell, leading to a cascade of inflammatory events, recruiting various different cell types along the way, leading to this destructive lesion.

It's interesting to note, as we have learned more and more about the effect of pharmacologic and biologic agents, as well as time, as we learned about the disease. Those drugs, nonsteroidal anti-inflammatory drugs that appear to affect prostaglandin synthesis way down here at the effector level don't seem to have the same kinds of side effects that the drugs that affect much higher in this cascade, those drugs that seem to affect cytokine production or cytokine interaction with various different cells, or even cell-cell interactions.

And these side effects, some of which

we're going to be talking about today, are inherent to the kinds of drugs we have available to us to treat rheumatoid arthritis. In fact, you heard many of them yesterday in the discussion of the TNF alpha inhibitors.

So what is the impact of rheumatoid arthritis on the health related quality of life? Well, there's clearly pain and suffering. There's decreased physical functioning, increased psychological distress, decreased social functioning, thus increased isolation, increased health care utilization, and thus increased costs, and increased work disability.

Our goals in treating this disease include halting progression of the disease, which is a word chosen quite specifically, that word "halt." It's something that we're still striving for, and despite some of the things that people have read about or heard about, we are not yet there. We do not have drugs that stop entirely the disease progression.

We maximize functional independence,

optimize the treatment of pain and inflammation. We obviously would try to enhance quality of life, particularly health related quality of life. We want to minimize the potential for toxicity, part of the discussion we'll have today, and provide easy access to care at reasonable cost, a clear indication of some of the problems that we have in developing new therapies.

I thought it would be interesting to see where we were 110 years ago. Basically this is extracted from the standard Textbook of Medicine in 1892, and most of us would agree who are M.D.s in the room that Sir William Osler is somebody that knew something about medicine.

And basically what he was referring to here is the treatment of arthritis, and the quote is, "Many cases are greatly helped by prolonged residence in southern Europe or Southern California. Rich patients should always be encouraged to winter in the south and in this way avoid cold, damp weather."

(Laughter.)

DR. SIMON: There clearly are reasons why one wants to be supportive and educate patients, but I'm not entirely sure that's the right way we should do that today.

Today we have a different series of options available to us in addition to education, support, exercise, and wintering in the south, which might be listed here. I actually have pointed out here that there is a -- and some people have pointed it out to me that this bullet is smaller than the rest, and I do that on purpose.

Nonsteroidal anti-inflammatory drugs and selective Cox-2 inhibitors, despite my background, are not drugs that do anything but palliate pain and inflammation, particularly in this disease. The drugs that are really important for this disease are those with the bigger bullets, and they include disease modifying anti-rheumatic drugs, immunosuppressives, glucocorticoids, biologic agents, and some of the investigational agents that we know about, but you guys don't yet know about, but it's pretty cool.

I thought it would be useful to look at before 1985 and then move up to be able to see where we're at. So these are the drugs that were used prior to 1985 or so. I'm not being incredibly accurate about this, but '85 is about the right time.

There were anti-malarials, IM gold, penicillamines, cyclosporins, azathioprine, cyclophosphamide, and chlorambucil. There are, I'm sure, people in the room that would say, "Geez, you would never use such a therapy for this particular disease," and they might choose chlorambucil or they might choose cyclosporins. They might choose azathioprine. But remember where we were at in 1985. Now, many of these drugs were not actually studied specifically for the disease rheumatoid arthritis.

Now, for many years it was considered standard of care to be cautious and not expose patients to potentially toxic therapy which had not clearly been shown or demonstrated to have a major impact on the disease. In that time, and even

today, the diagnosis is clinically driven. There are no yet biologic markers that specifically diagnose the disease, and many early patients suffered likely viral arthritis and not true rheumatoid arthritis, and these spontaneous remissions were probably not true RA.

So we always believed that there was some segment of the population that would get better by just palliating their pain and giving them time to get better.

Thus, a treatment pyramid emphasized slow progression of therapy from least effective modalities, but maybe safer in general, to palliate the pain and suffering to potentially more effective, but also associated with more potential risk of adverse events.

So three choices that I made of that original list are shown here in yellow. So many may remember that the anti-malarial drugs were fortuitously discovered when during World War II they were given for anti-malarial prophylaxis or anti-malarial therapy, and those patients who

concomitantly had rheumatoid arthritis got better.

Now, they're a pretty safe drug. They're reasonably well tolerated, although the list of adverse events in the PDR is about two pages long. You can't even get it on a slide. The major toxicity is retinal toxicity, directly related to drug pigment in the retina leading to blindness.

IM gold, a standard of therapy for many, many years, requiring injections periodically. It had previously been used to treat infections. Patients concomitantly having rheumatoid arthritis sometimes got better.

In 1966, the Empire Rheumatism Council studied IM gold therapy for the first time in a rigorous way, demonstrating significant improvement, an occasional case of remission, and significant risks in over 40 percent of the patients with chronic use. Heavy metal induced kidney damage was recognized; bone marrow suppression; liver effects; skin; vasculitis. And yet for 30 years it was the mainstay gold therapy of our treatment of rheumatoid arthritis.

It's interesting to note that cyclophosphamide, probably one of the better therapies that we have to treat this disease, a anti-cellular therapy, it showed significant benefit in the few studies that were done. It decreased disease activity and clearly showed a robust X-ray benefit in the one study that had been looked at.

Unfortunately chronic oral therapy increased the risk of urogenital cancers, leukemia, immunosuppression, bone marrow failure, nausea, vomiting, and hair loss, not an inconsequential list of potential therapies. When I trained, this was the list of options that I had available to me that I would use, but times change.

Now, the known truths were that the nonsteroidals, as I mentioned, were palliative, and that DMARDs, the disease modifying drugs that I just listed, were important for those patients with progressive disease would likely take about six months to know whether there was any benefit or not, and it would likely take six months or eight months before we would start therapy.

So it was 14 months or so before we determined that someone would respond. They were potentially toxic. They were associated with significant risk. They required often weekly surveillance at the initiation of therapy and, if subsequently tolerated, would require monthly visits, requiring CBCs and various other tests to ascertain whether or not they were actually being safely used.

Many patients had not an adequate response or developed adverse events, and the standard of care was still associated with damage evident by X-ray and progressive loss of functional status even in patients that were responders.

So what happened after 1985? Well, one thing happened, which was methotrexate became popular to look at again. I say "again" because it was first studied in the 1960s, but people were concerned about the use of a, quote, unquote, chemotherapeutic agent in the treatment of a chronic, quote, unquote, non-fatal disease.

But as we know today, in fact, it does

shorten life span. It is a fatal disease.

In 1985, there was a new description of the use of a low dose form of methotrexate at 7.5 milligrams weekly, which showed some benefit in a tiny study. Subsequently the dose in the clinical practice has risen. Most people are using about 15 to 17.5 milligrams weekly, and it was clearly better tolerated than some earlier, previously used DMARDs. There was some evidence of true disease modification, slowing of X-ray progression, for example.

But the potential adverse events included progress liver disease even while the patient was consistently monitored; lung fibrosis; acute pulmonary disease; bone marrow suppression; and immunosuppression.

This slide shows an interesting observation in 1992 performed by Pincus and others, and I show this because this is what it was before, which was that very few patients actually stayed on any one of several therapies, hydroxychloroquine, penicillamine, parenteral gold, oral gold, or

azathioprine, for any period of time once they were started on therapy until methotrexate, when in fact clearly for the first time -- and that's this line here -- people started to stay on it longer.

As I mentioned, it was better tolerated than the other DMARDs that we had previously been using, and patients seemed to be performing better on it so they stayed on it for a while. So this was quite encouraging.

So we move on from 1985 to now, and you'll notice I've changed the title from DMARDs to DMARTs because disease modifying anti-rheumatic therapies are now available both from the biologic side and the drug side, and they include sulfasalazine, methotrexate, leflunomide, the biologic response modifiers inclusive of the TNF alpha inhibitors, as well as Interleukin-1 receptor antagonists.

What are the advantages of these therapies? They've been show in robust clinical trials in this era that they slow disease progression. They've been shown to sometimes in

some studies improve functional disability. They decrease pain. They interfere with their processes, and in so doing, they clearly have been shown to retard the development of joint erosions by X-ray progression.

So this slide shows those drugs that have been approved and the indications for which they've been approved based on the new RA guidance document of the late '90s. In yellow are the specific indications, and in white are the therapies.

And as you can appreciate, most are approved for the presence of signs and symptoms, and then several are proved for structural damage. Leflunomide has been approved. It's not so suggest that methotrexate or sulfasalazine have not shown in the same clinical trials similar kind of data. The problem is that nobody has actually invested enough to take some of these older therapies for getting an indication at this juncture.

More importantly, it also suggests something about how one reports toxicities with

these older therapies. Many people don't report toxicities with older therapies because we already know everything there is to know about them.

So likely in the same way we don't give indications, we don't hear about safety issues with some of the older therapies.

Now, I'd like to point out that major clinical response, complete clinical response and remission, no drug therapy has achieved that at this point in time. These are clearly delineated within the guidance document of how to achieve it, and nothing has achieved it yet, and I point out that infliximab, as mentioned yesterday, is the only therapy to date receiving a prevention -- not really a prevention of disability claim as per the label, but actually improving physical function.

So the following five slides show the ACR 20, 50 and 70 for each of the products considered to be the disease modifying therapies. I have extracted this data specifically from the FDA approved label. I wanted to do this because of what was mentioned this morning at the open public forum

about benefit of one therapy for another.

I want to say it's incredibly difficult to compare these data across clinical trials without head-to-head trials due to differences in trial design, patients recruited, activity of disease, prior therapies, length of time with the disease, and the "et cetera" probably includes 15 other reasons why we shouldn't be comparing across trials.

The reason I have five slides is that if I put it all on one slide, it would be trying to do that. I don't want you to think I'm suggesting that.

But with all of these caveats, all of these therapies have a similar benefit. It's expressed by the ACR 20 measure, and often this same benefit in some of the therapies requires combination therapy to achieve it, and that combination therapy is often expressed in relationship to the concomitant use of methotrexate.

So one of the few areas where we can actually talk about the comparisons of leflunomide, sulfasalazine and methotrexate are within actually

the pivotal trials for the approval of leflunomide for signs and symptoms.

I only really want you to look at the yellow column, which looks at the ACR 20 response rates for leflunomide compared to methotrexate, placebo or leflunomide and methotrexate. I'd like to point out as mentioned this morning, there are differences between the US301 trial and the MN302 trial. Those differences are extraordinarily important to understand.

Firstly, different patients were recruited in these trials. These patients in the European trial had shorter duration of disease, more active disease than the patients in the US301 trial. These patients had longer disease, more chronic disease.

Secondly, even more importantly, folic acid is a concomitant drug used in the United States in almost 100 percent of the patients who are treated with methotrexate, and in fact, in this trial it was well close to 100 percent of the patients on folic acid. In Europe they rarely use

folic acid.

It is well known that folic acid decreases the toxicities of methotrexate, including stomatitis, hair loss, and even LFT abnormalities. In the study in Europe no patients used methotrexate, but in that context, folic acid also decreases the efficacy of methotrexate.

So as you can see, methotrexate here at 65.2 percent and here 45.6 percent.

So let's go back to leflunomide and clearly see that leflunomide at 52.2 percent, 54.6 percent, and 51.1 percent, not an inconsequential benefit in terms of signs and symptoms.

Looking at etanercept, in fact, looking at the evidence as per the FDA approved label, in Study 1 here you can see at six months a 59 percent improvement; in Study 2 a 71 percent improvement with concomitant methotrexate; and in this study, which was the arthritis study, which, in fact, is the only one you heard about yesterday from the sponsor unfortunately, actually suggests a much higher response rate, but these are patients with

very early disease. These are patients that clearly could benefit from aggressive anti-inflammatory therapy who had not yet sustained significant damage.

In fact, most interesting about this is that if you look back at the methotrexate history of development, patients respond much better to methotrexate with very early disease and much less as the disease has progressed over time.

Then also if you look at infliximab, and I remind you that in this context all of these data are expressed as infliximab with methotrexate, not as an alone monotherapy, and as you can see, the ACR 20 responses range from 42 percent to 59 percent, depending on dose.

Now moving on to adalimumab, the most recently approved TNF alpha inhibitor as per the label, one can see at six months a 53 percent rate of improvement at 40 milligrams weekly and 46 percent every other week, and then in the context of use with methotrexate, 59 percent at month 12.

So therefore, the TNF alpha inhibitors

and Arava or leflunomide, methotrexate and sulfasalazine, all have very similar ACR 20 responses as monotherapy when studied, and even sometimes with combination therapy they are the same.

This is the one slide looking at the non-TNF alpha inhibitor biologic responder modifier, which was IL-1ra, or Kineret, and again, pointing out just in the yellow month six, which is at 100 milligrams per day at 38 percent response, and in this study here at month six a 43 percent response.

So all of these data led to a clear paradigm shift in our weird treatment pyramid. We realized that, in fact, conservative care in the patient who had real diagnosed rheumatoid arthritis was probably not a wise thing to do. Remember, physician, first do no harm.

And clearly we need to be more aggressive in our therapy. So the disease modifying anti-rheumatic therapies clearly improve patient outcomes by improving signs and symptoms, by decreasing pain and inflammation, and they were

clearly shown, although I have not shown this evidence, that they regarded X-ray progression.

Thus, the standard of care today is to start aggressive therapy as soon as a certain diagnosis of progressive disease has been made.

I will not show this slide, but Dr. Wolfe in the audience reminded me that he, in fact, has shown evidence that the length of time patients stay on these drugs today is very different than in the slide that I showed you from Pincus, where they rarely stayed on the drugs for a long period of time, and under these circumstances actually tolerate these drugs reasonably well.

But even so, there is still no cure. Real remissions are rare. Ideally we would prefer a robust ACR 50 and 70 response, not yet seen with any of the monotherapeutic interventions. The data from the clinical trials really only approximate what may happen in the real world. Is a one or two year data set reasonable to predict long term results over 20 or 30 years?

Most patients need access to many

possible therapies, since there is no way to predict who might respond to any one therapy. Thus, it's important to have available as many potential therapies as possible with an acceptable benefit to risk ratio.

I'd like to take two seconds and review the Arava regulatory history as we move into the rest of the agenda.

The original new drug application clinical program began in 1989, and the leflunomide clinical program consisted of the three randomized controlled trials that I showed you before on that slide about leflunomide.

The U.S. trial, which was US301, was designed as a two year study with a primary analysis for efficacy at one year, while the two other pivotal trials were one year and second extension years were added on which required new patient consent.

It was a unique design, which addressed the problem of placebo and it led to a short placebo exposure period at four months and then a subsequent

conversion to active therapy in all patients who were nonresponders.

This led to a significant problem in data analysis that you will hear about today.

The original NDA was submitted in February of 1998 and includes the proposed claim of improvement in signs and symptoms of rheumatoid arthritis with retarding of X-ray progression. it included the proposed claim of improved physical function or functional ability, reduced disability, and improved health related quality of life, and the agency at that time granted priority review based on need.

The Arthritis Advisory Committee, this august body of August 1998, concurrence with the FDA was shown that studies demonstrated benefit for signs and symptoms, as well as X-ray benefit. A question was raised: should leflunomide be approved for the prevention of disability to the committee?

Now, it turns out at the time that this was all happening the FDA was creating a guidance document for the treatment of rheumatoid arthritis,

and it turns out that that updated draft FDA guidance document came out in March of 1998, a month after the NDA was submitted.

This draft newly defined the claim of improvement in physical function and disability and required two to five years of data. The exact type of the study to achieve blinded two to five year data was undefined. Was it to be blinded? Was it be controlled? Was it to be randomized?

Exactly how that was going to happen was not defined within the guidance document.

The AAC, the Arthritis Advisory Committee thus gave an answer to that particular question. It gave a reasonably good preliminary consensus that the data set was reasonable. The new guidance, however, which required the two to five years of data suggested that the committee should not recommend action because there was not two years of data to be shown, and there was only one year of data at that time.

So the leflunomide NDA was approved in September of 1998 for the treatment of active

rheumatoid arthritis to reduce signs and symptoms and regard structural damage. The three studies were then ongoing, the two studies for extension and one study that was a two year study, all of which provided blinded 24 month data to support the prevention of disability indication, and the FDA guidance for rheumatoid arthritis products was finalized in February of 1999.

And again, just to remind you that this guidance required at least a two year study duration of a known type; a validated measure of physical function to be measured, either the HAQ or AIMS were suggested; a validated generic health related quality of life measure was also to be included as supportive and should not worse, and what was suggested was the SF-36.

But what's very important within the guidance document is that there was a requirement that you had to demonstrate improvement of the signs and symptoms first.

Then in 2002 a supplemental NDA was submitted from the sponsor describing improvement in

physical function after discussions with our division, and these discussions were associated with the approval of one of the biologic DMARDs based on one year blinded data with a second year follow-up of that data demonstrating durability of that response in those patients who were responders.

Now, it turns out there were a large number of the patients in the second year who were retained within the trial.

So in conclusion, we have reached a time period in the treatment of patients with rheumatoid arthritis where there are several different DMARTs, sulfasalazine, leflunomide, methotrexate, etanercept, infliximab, and adalimumab; that improvement in signs and symptoms expressed in terms of an ACR 20 responder index, these therapies have similar effects, with effect sizes ranging in the context of ACR 20 responses of about 26 to 45 percent, with the context of different trials, different patients, early versus late disease, how many other drugs the patients failed, other concomitant therapies, such as folic acid,

combination therapies, et cetera.

There is a clear, been proven delay in X-ray damage progression by about the same degree when measured, and that potential adverse effects, although of different types, are not uncommon with any of these therapies, and all convey certain risk and potential risk even with appropriate use.

So I'd like to move on back to, Mr. Chairman. Thank you very much.

CHAIRPERSON ABRAMSON: Thank you, Dr. Simon.

We have a couple of minutes if any of the panel members have a specific question for clarity from Dr. Simon.

(No response.)

CHAIRPERSON ABRAMSON: Okay. If not, then we'll move on to Dr. Fries to discuss the health assessment questionnaire.

DR. FRIES: Thank you, Steve, and I feel honored and very pleased to be discussing Big Sky issues with you for a few minutes today because back when this story became some 20-some years ago, and

some people that were involved with that are here in the room, we wouldn't have ever had this discussion because we were getting too far away from the quantitatable things and into the soft, wishy-washy things that patients reported and patients said and we were leaving science behind.

So I hope to convince you that this is no longer an appropriate view and that taking what patients really do care about and putting that first and foremost is part of a transition that we should have going on.

So I'll speak from the standpoint of the development of the HAQ, recognizing that Jennifer is here, who was involved in the beginning efforts of the AIMS instrument, and there are other people. Fred Wolfe is in the audience who also has had a great deal of experience in these areas and is widely cited in some of the background information which is provided.

The health assessment questionnaire was originally called the AAQ or the arthritis assessment questionnaire before it was recognized

that it really had much more in the way of generic characteristics than disease specific characteristics, and I'll return to that.

The publication of both the AIMS and the HAQ articles were in 1980 in Arthritis and Rheumatism in the same issue. The HAQ paper has become the most cited rheumatology article over this period of time.

The current paper, which is included in your handouts, which came out in January of '03 cites actually some 70 different languages that it has been translated in and also a variety of areas in which it has been used in clinical practice, particularly by Drs. Wolfe and Pincus and people that have worked with them.

ARAMIS itself which I direct, which is the arthritis, rheumatism and aging medical information system has administered well over 200,000 administrations of it. In terms of cited publications validating the instrument, they now number over 400. Most of the more recent ones are cited in the Journal of Rheumatology article that

you have. It's been used in a lot of disease areas in studying human aging, particularly in musculoskeletal aging, in AIDS, in arthritis, in connective tissue diseases and basically all of the rheumatic diseases with minor modifications.

It's not quite a required disability outcome variable for clinical trials, but it and similar instruments have been mandated in the ACR list and the OMERACT lists, and one of the questions that perhaps will come up today is how should you actually compute something like an ACR 20. Should all of the potential ingredients be used? Can different people pick and choose from different areas as to which ones they want to count? How do we level the playing field? Are we having the most important variables required for the ACR 20 or are we not?

So some of these issues, I think, are really important and some of our greatest fans have made the argument that, in fact, the HAQ disability index is the dominant outcome variable in clinical trials in rheumatoid arthritis, and why should we

have anything else?

That is not my position, just to clarify that right off the bat.

(Laughter.)

DR. FRIES: But it is when I look at studies the first thing I look at, is the HAQ disability index, and then after that I look at the ACR 20 and all of these other things to see what it is, and so there are definitely some issues around this point.

Now, I've got to introduce this by saying that this is a paradigm shift that we're talking about that wouldn't have been present 20 years ago. It's a processed outcome change from process variables that a patient doesn't feel or perceive to outcome variables which are very central to their way of living.

It's a move, as you heard from Lee's discussion, from short-term outcomes to long-term outcomes, and we still continue to have this tension between what is long enough in a 25 year disease. Is two years long enough, five years long enough,

ten years?

What are the questions of sequencing? How do we handle the integration of new drugs as they're approved into the sequence of difficult clinical decisions that we have?

So as we begin to think about diseases as 25 years in length, we clearly have to move our studies. Our studies have to move from cross-sectional snapshots to longitudinal studies of the same patients. We'd like to move in a sense from the mastery of the physician to the mastery of the patient, to the self-management to the individual decision making, the autonomy expression that the patient can have to the greatest degree that is possible and consistent with best results.

And clearly, this takes us to the oft recommended or argued partnership between patient and physician. Clearly, you have to apply science and the best science to these decisions, and clearly the patient has to put their values into the mix and determine what, in fact, is better from point effects to cumulative. There are several ways to

get from being normally functioning to being severely impaired.

One way has you maintaining your function for a long period of time. Progression has been halted, as Lee would put it, or postponed, as I would put it, and you may still get there, but you postpone this getting worse.

You also could have something which deteriorates very rapidly and essentially stabilizes at a very low level of quality of life. Those might have the same point endpoint, but they'd have very different cumulative area under the curve endpoints.

So we're tending to move toward cumulative endpoints and toward area under the curve endpoints, and I'd submit that this merits the term "paradigm shift." I think it really does. As Lee emphasized, we changed abruptly, exactly opposite our general approach to rheumatoid arthritis because we really couldn't let people get crippled before we treated them. That will be expressed by any rheumatologists who are in this discussion today.

The world is very, very different with

the newer drugs and the newer philosophy of approaching them.

We have a paper currently in press which demonstrates and documents in our data sets a decrease of about a third or more in cumulative disability in patients with rheumatoid arthritis over the past 20 years, and the concurrent changes are those which we've been talking about.

So data is beginning to come out that not only are these theoretical shifts in paradigm. They are real changes in real people over this period of time, and they are substantial advances.

Now, I'd like to try, and this is the big sky stuff. So if you don't mind being in church for a little while.

Plato described ideals of things, and so we call platonic outcomes, and this is sort of the basic idea when you start talking about outcomes and patient oriented outcomes. You sort of have to get back to ground zero and figure out what are the first principles. What are the things that patients want and how do we redirect our medical care system

to get patients to the kinds of things that they really and truly value.

Plato's values were universal. In this instance, we wanted to emphasize patient directed. I want to mention disease independent because we've had this, I think, rather non-helpful distinction between generic and arthritis related measures. This has been in many ways a false dichotomy because some of the most widely used instruments in other fields of medicine happen to be developed by rheumatologists and then exported into other areas.

So they kept this. These are like disease specific. As I indicated, the HAQ has been used widely in human aging and many, many different disease areas, and I would hold that you have to have or it's a strong desirability to have instruments which are disease specific or almost disease specific.

You'd like to figure out what domains or dimensions you have, and you'd like to ideally make them mutually exclusive and collectively exhaustive so that you've got the whole universe, what patients

might like you to do, with it included, and yet you have separate numbers that aren't too many so that you can actually compare things. So you'd like to have things that are mutually exclusive and collectively exhaustive.

Now, it turns that only generic measures can be platonic, that can approach this kind of ideal. Otherwise we get ourselves into a linguistic bind in which we have an entity we may term "disability" or something else, and we consider the disability as one thing in aging people and another thing in sclerodermal people, and another thing in rheumatoid arthritis people.

No, disability has got to be disability. It's a universal concept, and diseases may affect things more or less with it, but somehow or other, these concepts are not different across diseases. It's the diseases that differ in their quantity of each of the problems.

So we have generic instruments. We have disease specific instruments, and probably -- and we would argue now that we should be moving toward

using one of a small number of generic instruments with disease specific supplementation in other areas. You have to be able to examine effects in diseases across diseases, which means the same measure. You have to recognize that one size doesn't always fit all, and there needs to be the ability to have supplemental questions in particular areas.

Now, our concept, and it follows in many ways Kerr White of Hopkins now four decades ago sort surveying what it is that patients really want and kind of coming down with what we have advertised as the five Ds of death, disability, discomfort, drug problem, drug and doctor problems, and dollar costs.

And those are essentially the dimensions that people will select if given the options to check. If you don't give them a menu, they won't put economic in, and they often won't put iatrogenic in if they're doing it free form.

But if you ask them to actually list, then they say: I'd like to be alive as long as possible. I'd like to be functioning freely and

normally. I don't want to hurt. I don't want any side effects, and I want to remain solvent in a difficult world.

So that's what patients say, and they're not quite, if you analyze them, mutually exclusive. They're probably not quite collectively exhaustive, but there's an attempt to try and get this kind of an umbrella.

If one does that, then there are some automatics or subdimensions that you can consider under this, and then there are components, and so somehow as you worked out the components, you begin to kind of sum it up.

And we felt that one has more trouble in terms of defining this in quantitative terms than you do by defining each of these dimensions, where one can roll up data from a level to give you data at this dimensional level, but you have a problem with some uncomfortable transfers between death and dollars and things of that kind, if, in fact, you roll it up the last step.

So we've argued that a complete outcome

assessment program is essentially the full HAQ with its protocols, which measures each of these.

Now, I was asked to speak a little bit about disability and physical function and what in the heck we should call this thing that we all sort of know what the ideal is of it. Here are just several instruments. There are many different instruments, and of interest with the instruments is the McMaster health index questionnaire as physical function, social function, and emotional function.

That's kind of nice. It's a paradigm that has all of the domains. It's mutually exclusive, collectively pretty exhaustive. It's a nice, simple, logical frame. It has a thing it calls physical function, Nottingham health profile. It has a thing it calls physical mobility, quality of well-being; a thing that has an area called mobility and another one called physical activity; the sickness impact profile. It has a variety of things which involve physical and then a variety of things that involve other things.

You can see more or less sense in these

domains that different people have chosen when developing their instruments, but they all have included this entity of physical activity disability, although they've called it sometimes different terms in the subscales, but they mean the same thing.

How well is the patient functioning in sort of a positive sense? How disabled are they in sort of a negative sense?

And I felt that it's important to go back and look at the way in which the makers of an instrument have sort of categorized illness because you can find both the similarities. I showed you the HAQ before, the five dimensions of the HAQ, and you can find differences and you can find omissions and you can find duplications.

This is the HAQ. I show you in four slides sort of the two page HAQ here. Date; the term "arthritis," which in generic representations becomes considering all of your health or considering all of your scleroderma is.

So there's an area in which disease

specificity comes in with this word in the stem. That's the only place. All the rest is generic, and it happens because we'd really like to separate out comorbidity coming from other places if we could, and so this is just an attempt to say, okay, we're looking at arthritis related disability.

This is the way the questions go. A dressing and grooming category; are you able to dress yourself, including tying shoe laces and doing buttons, shampoo your hair? Without any difficulty, with some difficulty, with much difficulty, unable to do. Scored zero, one, two, three.

The highest of each item in each category is selected so that if one can check here and here it would go in as a two. I'll show you the way in which aids and devices are done. This is to increase the sensitivity of the instrument because, in fact, patients move slightly irregularly through different kinds of problems, and it's nice to be able to pick up the most sensitive disability while having one necessary activity of daily living included.

The intellectual heritage really comes from the Steinbacher criteria, the ARA functional class, which is in -- somebody could help me maybe -- 1942. It's a long, long time ago, and it had the same concept. There was Class 1, 2, 3 and 4, which were conceived just like this.

It was far too crude in its specification, but it was used to classify people with rheumatoid arthritis and other forms of arthritis, and that was the ARA, old American Rheumatism Association functional class, and that's what it has.

Then there's our other categories, such as arising, eating, walking, and then there's an aids and devices section, and this is required to clarify the ambiguity that arises when somebody says, "Hey, I'm walking with some difficulty, but I'm using a cane," or a walker because we would really like them to have a higher number if that's the case.

So they check the devices that they're using, and these tie back and will take people to a

score of two even if the patient hadn't said two in an area where they're using an aid or a device.

This, again, increases the sensitivity and gets us to the issue that we're really interested in, which is not the effectiveness of aids and devices, although if we want to do that we can just score it without this section, but it's how disabled the patient is or what is their level of physical function.

Now, hygiene, reach, grip, and so forth.

Now, of interest, and I don't really think it's relevant to today's thing, but the study which was reported as nearly as I can tell from the background materials didn't use the HAQ. That is the story that included leflunomide and methotrexate. It used a combination of the PET and the HAQ, which is pretty awful.

I hope what you're able to see here is this is cleanliness. Okay? It's simplicity. It's clarity, and we've been over every word, every place, and we've looked at the display techniques and so forth, and you do that in order to get

maximum comprehension across educational level groups. You'd like it really to be crystal clear.

If it looks like it's so simple it was done on the back of an envelope, that's perfect. You know, the idea of making it really complex -- and if you look in the briefing materials where they combined the PET and the HAQ, it triples the length of everything, and it makes it really quite confusing, and I think it may have carried the PET along, and it may have lost a little bit of the HAQ at least as designed, but it still worked. It still worked fine, and we had, again, up toward the optimal performance of any of the measures that you use for measuring rheumatoid arthritis.

And then the pain scale, which is another of the ACR 20 criteria: no pain, severe pain, doubly anchored, horizontal visual analogue scale rated from zero to 100, and that's the short HAQ.

The long HAQ -- it's two pages, and it's scanned and works very, very nicely. The long HAQ is about 16 or 17 pages and deals with the economic

impact of disease, the side effects and so forth, and then they're associated with protocols that involve auditing of hospitalizations and auditing of deaths and use of the national death index and so forth, all of which go beyond today. I was just talking really about assessing functional ability and activity today.

Well, this is sort of a question that has been raised, and I guess the group can decide today. I've indicated to the FDA that I'm rather neutral toward what terminology is specifically used to describe this entity which we know what we're talking about. It should be of maximum clarity.

It's been pointed out that the term "disability" has a whole variety of other meanings, which could be confused with each other, you know. Whether you can get a blue parking sticker or not, and as disability, whether or not you can get certain kinds of payments from the social support system. This is disability, and as "disability" is used there, it's important to note that it's always a threshold phenomenon. You either have disability

and you can get the blue sticker or you don't have disability and you can't, and there are criteria and wars and fights about how exactly you should define that threshold, and that's because that's really wrong, isn't it?

I mean, disability is on a continuum or functional ability is on a continuum. It isn't like all or none that's there.

So one thing would be to say what's been done throughout the briefing document and what we always do is we say HAQ-DI. We don't talk about disability by itself. We talk about a disability index, which is a different kind of an entity.

So there's part of me that kind of prefers disability index as a term. Probably disability itself has more disadvantages than advantages, and we should probably perhaps move from that.

All outcome instruments that I've shown, they have a disability domain, but they often name it differently. The concept is the important advance, and that's what I'm trying to say here, is

it's time to get to this subject area and really enshrine it and make it one of the treatment goals. That's what I see the advance of, and I'll be happy with anything that you come out with that takes us in that direction.

These are the different things disability could mean, receiving payments, getting blue parking stickers, and so forth, several legal meanings, and then there are a variety of things that have been used that would be functional status, going back to the Steinbacher criteria, physical function, physical activity. Any of those things can be done.

There are some implications that have to do with are you inverting the scale and causing confusion. Should you go from three to zero or zero to three, depending on whether you call it physical function or disability.

We said there are like 400 articles out there, and they've all used it one way, and I sort of wish we had done it a little differently when we had started it, but now it's so enmeshed that one

would sort of like to continue zero to three HAQ-DI scores in because you know which way is up, and people have gotten used to that phenomenon.

HAQ or MHAQ. Now, we could generalize to other kinds of things. The HAQ and the MHAQ, which is a derivative instrument, uses the same eight categories, but I would hold that this group should be very aware because of the implications of decisions at the FDA level and the cost of clinical trials that sensitivity change is really the thing that one wants in a physical function variable because greater sensitivity means greater power. Greater power means fewer patients. Fewer patients means lower costs.

So one can actually vary the cost of a study very greatly by using instruments which are as sensitive as possible to change.

The HAQ's greater sensitivity which has been shown a lot of times is because of the additional variables. As I showed the highest score per category and the aids and devices adjustments, those are important features with regard to

increasing the sensitivity of an instrument.

Signs and symptoms. It has been posed to me. The question is: are signs and symptoms -- is disability or physical function a symptom in some way? Because some of the outcome variables like pain are.

I would hold that it's not. It's an aggregated outcome dimension, conceptually different from medical process, and it's a separate clinical indication, perhaps the most important. It should be a required measure for demonstration of efficacy, NRA, and there are several ways in which this could be done.

It won't be my decision. It will be our decision perhaps as to how, requiring all of the ACR 20 components to be used, using the same criteria for everybody, separating physical function from the others and making it a required one, sort of like an ANA and lupus kind of phenomenon. You have to demonstrate improvement in physical function and some other list of things.

But, again, I think the principle of

using the same criteria for all studies does make a certain amount of sense with regard to approval of drugs' NRA.

What duration I was also asked to kind of say. The placebo control issue, I guess, will be the subject of a lot of discussion here today. It's not at all surprising. It may be surprising to see it, but it's not surprising those who take our patients that patients with rheumatoid arthritis don't do real well on placebos, and they tend to drop off and they tend to demand to leave studies in large numbers.

And actually on the ethical ground they're destroying their joints and they're getting irreversible changes in physical function and other kinds of things to happen.

So placebo groups will drop out, and they'll drop out rather rapidly, and it creates a methodologic dilemma because we'd all love to see truly long-term placebo controlled studies so that we had something rock hard to compare it with, but we ain't going to see that because the people that

drop out are not the same as the people that stay.

So you have the preferential dropout of the sicker patients, and that gives you a problem, and it's a cross-over problem, and there are some ethical problems, practical problems associated with it, and it looks like you probably, to me, that you can have a shorter placebo period, perhaps figure it out, but I doubt if it's really going to very often go beyond 12 months without getting into trouble.

It also can be a lower sample. It doesn't necessarily have to be as many people in it as you have in, let's say, your two comparator arms.

For your active comparator, you have the same cross-over problems as -- but they just happen a little bit later because people change drugs, too, and they drift off because they're not doing as well as they thought they ought to on this drug, and so they drift off.

So if you start talking about three, four, five year studies, then you really can't get enough people staying in the active comparator group to be really useful either.

And our people have been talking a little bit back and forth about how long you stay on different drugs, and this in our experience is a real changing phenomenon. The more alternatives there are -- some of the neglected reasons for changing drugs is a new drug comes on the market and so you have more options.

So we're seeing a real decrease in methotrexate length. We have people who are not staying on it for five or six years. There are too many other things that you could put people on and be happy with. So those numbers are actually shrinking down

What is increasing and continuing to increase is the percent of disease course on a DMARD or a DMART. I hate to change these things, Lee.

(Laughter.)

DR. FRIES: So anyway, that's a problem.

And then there's this neglected kind of thing that says that, hey, there are other things that affect functional ability in patients with rheumatoid arthritis, and it may be congestive heart

failure or it may be because you're 93, but at any rate, some of these things begin to after some length of time blur our ability to separate out the rheumatoid arthritis as a cause of loss of functional ability and the disease itself or the other parts of the life.

Now, all right. This is, in a sense, the key answer to a lot of the questions that we have. This is what we call a therapeutic segment. This particular one is methotrexate, and this was published in JRHEUM last year. This is looking out over 84 months of treatment at patients who were on the drug for different periods of time, and these are looking at their HAQ scores.

In the real world, these numbers are not as big as the ones that Lee showed you. They go down from 1.5 to 1.2 on average. The lowest area of functional ability or disability where it's at its lowest is actually out about 36 months into treatment. So there's continued treatment through the earliest part of what we would call the therapeutic segment.

Then there's a plateau period, and then there's a decline in which the disease progression overpowers that particular drug in individuals, even out here with people who are self-selected for having done reasonably well on methotrexate.

So one sees this, and it's quite reasonable to say that this ia a general figure, although we haven't yet looked at leflunomide and some other drugs, but I think as clinicians we would not be surprised that there is a period of biologic effect, a period of consolidation, and then a period in which the disease reprogresses, begins its reprogression.

And as we think strategically a lot of what we need to do is to figure out at what time you jump ship. You know, some place down here perhaps you go to the new drug even though the patient is doing reasonably well in anticipation that something else is numbered.

So we're thinking a lot about how we would strategize these things so as to fill up a 25 year course, anticipating that a lot of other drugs

would be coming on as time went along.

So it's reasonable, I think, to expect that any of the TNF alpha drugs or leflunomide or other drugs coming on will probably show something like this, and then as Lee showed, the decreases that we see are actually fairly similar between these drugs. The TNFL for drugs seem to be adding a release of toxicity feeling, a gestalt in patients as much as they actually change.

Because it looks as though, for example, methotrexate plus leflunomide would give you, if started simultaneously, would probably give you similar amounts of drop that one would get from one of the TNF drugs, but all of those drugs are probably going to do something like this, and so the question is then how long a study is necessary.

I mean, it would be a question of are you concerned that a drug which does this in the first 12 months is now all of a sudden going to go up, you know, in the second 12 months; its effectiveness is just limited to some kind of period of time.

I don't think so. I don't think we have any indication that drugs lose their effectiveness per se. We do have some evidence that the body grows weaker, and the disease may be accumulating slower progression over a period of time, but I really think that one can predict the fact that you have had an improvement in functional ability on the basis of the initial drop.

So, in my interest as you would have perceived in changing the paradigm, it is saying that let's have randomized trials of whatever period of time. Clearly they won't be less than a year in the initial ones, and then have a follow-on period with the same patients or with other patients, hopefully with common protocols across drugs so we really can get some kind of an early warning system.

We have our protocols. Some other people have theirs, but we should be doing the same protocols across different drugs so that we can begin to get even if it's in the observational setting some direct head-to-head comparisons, and we really need to do that, you know.

And if the same databases can survive all drugs, this is like you can identify the protocols and each company could execute the same protocol, but that wouldn't satisfy us as much probably as if some sets of databases studied in parallel all drugs and used their own comparisons with their own people and their own scoring and so forth.

So I see this as more important than the length of time. Now, this is where I'm perhaps going farther than the group wants to go today, but who should get the new indication?

I mean, I hope I've made an argument we should have an indication in this area. This is a very important area. Okay? And this would be my personal conclusion that fits a lot of the data that you have. Are the sponsor's data sufficient to document improvement in physical function or whatever we want to talk about that? I think that's clear.

So are the data of several other sponsors. See, of interest once you've gone into

the ACR 20, 50, 70 kind of game, you've already got HAQs for however long these studies were. A year? You know, even though they weren't reported out that way, those data exist in all of these areas. They've been reviewed by this committee and by the FDA and agreed that they are high quality and so forth, and so there are several other sponsors who can really make a similar type of claim, I think, and to my mind they don't have to do new work to do this.

If, in fact, they've already met the same criteria, they should be able to file that area. Much of the data has already been reviewed by the FDA, and so I close with this.

Why not, if we're going to move toward this, open the doors for this indication? It's an important indication, and it would be nice to have a number of drugs which had it.

Thank you.

CHAIRPERSON ABRAMSON: Thank you, Dr. Fries.

We have a few moments if members of the

committee have any questions for clarification on Dr. Fries' presentation.

Dr. Gibofsky.

DR. GIBOFSKY: Jim, I very much enjoyed your presentation of the five domains, the five Ds. Can you help me get a handle on to what extent patients weight those five Ds in trying to make assessments about their therapeutic decisions?

And as a corollary to that, to what extent should we be weighting those five Ds in assessing claims for indications and benefit-risk ratios?

DR. FRIES: Yeah. Well, with the caveat that studies designed different ways have come up with different things, if you use the patient global where you have an analogue scale and, you know, it says, "Considering all of the ways your arthritis affects you, mark your score how well you're doing on a zero to 100 score," and use that as a gold standard, then you find in rheumatoid arthritis that there's about two times -- it basically turns out to be disability and pain that they rank again in a

free form area, and it's about two disabilities for one pain.

In osteoarthritis, it tends to be the reverse with pain valued more as a determinant of patient global.

Now, patient global, as I indicated, all of these problems with kind of estimating a global entity because you're asking such a totally different question than when you're actually asking, let's say, a question in disability or functional ability, and there a good question is one that says, "Can you reach up above your shoulder and take down a five pound bag of sugar? Can you reach down and pick up a piece of clothing from the floor?"

These are very, very precise things, and if you say, "How are you doing, you know, with your arthritis?" you get a very different response. A lot of people say, you know, "I have my faith and every day is a blessing to me. I'm doing wonderfully," and then you have the opposite type of people who are always doing poorly, and it doesn't necessarily correlate with the harder notions.

So with that caveat, if you ask the question in certain ways, you can get people to be concerned about the cost of drugs to a greater degree or to have greater amounts of fear about the side effects.

So they are all sort of essential, and you can think of circumstances and patients in whom each of them would be dominant.

CHAIRPERSON ABRAMSON: Jim, a question over here. In distinguishing disability index and physical function, I'm curious about the HAQ. What are the domains that contribute to the disability index and how do they differ from other assessments of physical function? What is the Venn diagram like in that respect?

DR. FRIES: Well, there are the eight categories which I showed, and they are basically activities of daily living. They include both IADL, that is, instrumental activities of daily living, and ordinary ADL, a distinction that I haven't particularly found to be a useful one, but things like running errands and full daily activity are

called instrumental activities.

Anyway, where something like walking is a basic area, but the actual way in which the questions were drawn was that we took all of the questions that had been considered in ADL assessments prior to the HAQ, and we found 68 definable questions.

We did a big thing with all of the questions on everybody, and then we did correlations with an early HAQ, which was the mean of 68 questions. We looked for things which were redundant to others, questions like all of the walking questions sort of crossed over with each other pretty much, and then we looked at things that were correlated or not correlated with the overall index as being nondimensional, and then we collapsed the group down.

We started losing stuff at 20 collapsed into eight. Originally we actually had a sexual function question, which we removed because it didn't add anything to the accuracy, and it did decrease the percentage of people who completed the

questionnaire or completed that question.

So, I mean, that's the way it was derived. As you do that, you're carrying sort of the ghosts of questions which are not included in the final product. You see, they were included in the original 68, but not in the final 20, but that was because they redundant or correlated highly.

So in a sense you carry some of the meaning that was connoted by the entire data set. So it's pretty complete. If you want to do -- I mean, just to be fair and talk about limitations, the HAQ accidentally or deliberately picks up mental function, too. I mean, depression affects scores, for example, on the HAQ.

There are no questions about hearing or seeing or balancing your checkbook, and these are functional questions. And so something that was -- it's why I kind of waffled a little bit on the exhaustive nature of things. They're things that we don't have and most other instruments, as you look at the content analysis have the same kinds of problems.

So for certain things, we will add mental function areas and organs of special sites because they do contribute to function.

CHAIRPERSON ABRAMSON: Dr. Goldkind.

DR. GOLDKIND: Yes. To follow up that answer, what is the correlation between, let's say, a strict analgesic or a mood altering drug and a HAQ? Has that ever been looked at, simply teasing out --

DR. FRIES: An interesting question. Yeah, it's an interesting question as to whether you could use a tricyclic or something like that and change a HAQ score.

Fred, do you know of any such studies, looking at a psychoactive drug affecting HAQ disability index scores?

DR. FREDERICK WOLFE: No, I don't think there have been very many. It's a study that needs to be done, but I don't think it has actually been done.

DR. FRIES: Yeah, that's my same answer.

DR. SIMON: But, Jim, what about the

context of nonsteroidal anti-inflammatory drugs or simple analgesics? Do you believe that whatever is measurable within the context of improvement in the HAQ by such an agent, which actually has no fundamental benefit other than pain relief -- where do you see that in the context of what we're measuring?

DR. FRIES: Yeah, I think that's an important point. NSAIDs don't move HAQ disability index scores. They just don't move them. Three, six, nine, 15 months later they're just where they were. Sometimes things get a little bit worse.

If you take a look -- and the same thing goes for pain scores. Analogue pain scores do not get moved by nonsteroidals even though those are analgesics. Pain scores do get moved by DMARDs greatly and disability index scores get moved greatly by DMARDs, but I consider that, in a sense, an off-side validation of the studies, that in fact, they act like we would like them to act.

CHAIRPERSON ABRAMSON: Thank you very much, Jim.

We're going to move along because we're a bit ahead of time and are going to go directly to the presentation by the Aventis company and Dr. Rozycki will lead off.

DR. ROZYCKI: Good morning, ladies and gentlemen. I'm Mike Rozycki, from Aventis' U.S. regulatory affairs organization, and on behalf of Aventis, I wanted to thank you for the opportunity of being here this morning to discuss Arava.

By way of orienting our discussion this morning, I wanted to revisit the questions that will be considered by the committee this morning.

Does the term "physical function" or "disability" better capture clinically relevant information ascertained in the HAQ?

What duration of superiority study is needed to robustly identify improvement for disability and physical function?

The data that are needed to assess durability of effect beyond an initial superiority study period.

And then, finally, are the data on

leflunomide adequately robust to support labeling for improvement in physical function?

So this morning we're here to discuss the addition of a claim for improved physical function to the label for Arava. I wanted to just review what the treatment goals for Arava or leflunomide have been during the course of its clinical development; improvement in signs and symptoms of the disease; reduction of structural damage evidence by radiographic evaluation or erosions and joint space narrowing. These two items are already in the label.

And then what we're here to discuss this morning is improvement in physical function as measured through health related quality of life instruments, using specific measures such as the health assessment questionnaire for use as a primary endpoint and the more general measures, such as the Short Form 36 to capture the full effect of rheumatoid arthritis on the patient.

Now, Dr. Simon has reviewed the regulatory history of Arava already, and that makes

my job this morning a lot easier. There are a couple of points from the regulatory history that I wanted to review because they are going to be recurrent themes.

The first is that the NDA -- and I think my voice is probably going in and out on the microphone here -- the original NDA for leflunomide, which was submitted in March 1998 consisted of six or 12 month pivotal data from the three randomized controlled trials described by Dr. Simon, and the words that should be on this slide are "ITT cohort." This pivotal data constitutes the ITT cohort that we will be referring to in later sections of our presentation.

And then, of course, as Dr. Simon mentioned, the Arthritis Advisory Committee met in August of 1998 to discuss the claim for physical function, but decided not to vote because at that time two year data were not available for leflunomide, and of course, the leflunomide NDA was approved in September 1998.

Since the original approval of the NDA,

the three clinical trials that provided the original pivotal data were continued or extended, depending on which trial was involved and provided blinded 24 month data in support of the physical function indication as defined in the 1999 FDA guidance.

And, again, to revisit the study design, US301 was a 24 month study with prespecified data analyses at 12 and 24 months, and supporting data comes from the international studies, MN301, 303, 305, which was a six month initial study followed by six and 12 month extensions, respectively, and MN302/304, which was a 12 month initial study followed by a 12 month extension.

And to remind the committee that what we're here to request from the FDA is the addition of improvement of physical function to the current label for Arava.

Before we go on with the main presentations, we did want to acknowledge a large number of outside expert consultants that have been involved during the course of the clinical development of leflunomide, and many of them are

here with us to facilitate our discussion today. I won't read every name, but you can scan through the list of names that are up here.

So to continue on with the main portion of our presentation today, we will have a discussion by Mr. Joseph Doyle, who is with Aventis' Health Economics and Outcomes Research Group at Aventis Pharmaceuticals. He will describe how the methodologies for measuring physical function described by Dr. Fries just now were applied to the design of the three randomized controlled trials.

Mr. Doyle.

MR. DOYLE: Thank you, Dr. Rozycki.

Members of the panel, ladies and gentlemen, I recognize that a number of disciplines are represented today on the panel. So first allow me to review the patient reported outcomes, but physical function and health related quality of life that were included in the three leflunomide pivotal, randomized, controlled trials.

These patient reported outcomes include the health assessment questionnaire, commonly

referred to as the HAQ, the SF-36, or the Short Form 36, and the problem elicitation technique, or PET Top 5.

I will then review the relationship between treatment associated improvements in physical function as measured by the HAQ and the broader concept of health related quality of life as measured by the SF-36.

I will conclude with a very brief review of some terminology that will be used through the presentation today, such as the minimum clinically important difference, or MCID, and the number needed to treat, or NNT.

This terminology will be used by both Karen Simpson and Dr. Vibeke Strand in the presentation of the physical function and health related quality of life data.

We know that impairment in performance and physical activities due to active rheumatoid arthritis has significant effects on day-to-day activities and physical function, as well as health related quality of life. Inability to perform

activities of daily living occur very early into the disease, with 50 percent of the patients unable to work or work in the home within five to ten years of the onset of disease.

Measures of physical function, such as the health assessment questionnaire, are able to predict work disability as well as joint replacement and premature mortality.

Symptom improvement, as reported by the patients, has frequently been the only means of detecting treatment effects, and patient reported measures have always been a fundamental part of the drug development process.

When we talk about a chronic debilitating disease, such as rheumatoid arthritis, patient reported outcomes, such as physical function and health related quality of life are central in determining treatment effects and have become a focus of the drug development process.

Briefly I'd like to review the patient reported outcomes that were included in our clinical trial program. The HAQ as described in depth this

morning by Dr. Fries, is one component of the ACR response criteria. It is a valid instrument, widely accepted, and used in rheumatoid controlled trials.

I won't go into the detail on this slide as they were provided by Dr. Fries this morning. However, I'd like to mention that this is one item that I look for as well when I review rheumatoid trials.

The HAQ was included in all Phase 3 trials for leflunomide.

The HAQ is scored from zero, indicating no impairment, to three, indicating inability to perform activities of daily living independently. An increase of one unit per year over the first two years of disease results in a 90 percent greater disability over the next three years.

As demonstrated here, the HAQ DI score worsens and as annual medical direct costs increase dramatically.

In a meta analysis published by Scott, et al., examining longitudinal studies from the U.S., Australia, and the U.K. with standard care and

conducted prior to the introduction of newer DMARD therapies, it was found that by 12 to 18 years of disease duration that 50 to 60 percent of the patients with RA were unable to work or perform activities of daily living.

Until relatively recently, for patients with RA, it was thought that progressive loss of function was inevitable over time with standard care, including DMARDs and nonsteroidals. Even observational studies published as recently as 2000, reflecting more aggressive treatment prior to the introduction of new DMARDs, showed that stabilization of HAQ DI scores was the most that could be expected.

In contrast, recent randomized controlled trials in rheumatoid patients entering a second year of therapy utilizing new DMARD therapies as illustrated here with infliximab in the ATTRACT study, the HAQ DI is responsive and able to detect changes in physical function over time.

Improvements in physical function are seen at six and 12 months, and maintenance of this

effect is seen over 24 months of therapy. Based on this data, infliximab received an indication for improvement in physical function which we are seeking today for leflunomide.

A similar pattern of improvement in physical function and maintenance of effect is seen with etanercept over 24 months in the ERA study.

And as you will see again later today in the presentation, this same pattern of improvement in physical function at six and 12 months and maintenance of physical function over 24 months is seen in all three leflunomide clinical studies.

Another measure of patient reported outcomes in the problem elicitation technique or PET Top 5. The PET Top 5 asked patients which physical activities queried in the HAQ are most affected by their disease and that they most want to see improved.

And finally, the third patient reported outcome included in our trial is the Short Form 36. The SF-36, developed by Dr. John Ware, is the most widely used validated generic measure of health

related quality of life. It consists of 36 questions which are divided into eight domains, scored from zero, the worst possible score, to 100, the best possible score.

In addition, two component summary scores can be calculated, the physical component summary score, PCS, and mental component summary score, or MCS.

The SF-36 has been used in more than 200 peer reviewed studies of arthritis, and it was included in more than 30 randomized controlled trials for rheumatoid arthritis.

Originally, the SF-36 was not believed to be sensitive to change in RA. However, the leflunomide US301 study was the first study to show treatment associated improvements in health related quality of life in patients with RA.

Note that the SF-36 was not included in the leflunomide European MN studies, which were designed in 1993 and initiated in 1994, since valid translations were not available for many countries at that time.

In the leflunomide clinical study US301, and as expected in an RA population, baseline SF-36 scores prior to treatment, illustrated here in the lighter bars, show marked decrements in all domains of health related quality of life when compared to age and gender adjusted U.S. norms. These decrements are most evident in physical function, role physical, bodily pain, and vitality, but also general health perception, social function, role emotion, and mental health, hence, indicating the impact of RA on health related quality of life.

The physical component, or PCS, and mental component, MCS, summary scores of the SF-36 are calculated based on all eight domain scores. When scoring the PCS, the four physical domains are given the highest weight, illustrated here in red.

These component summary scores are standardized, using U.S. normative data to have a mean of 50 and a standard deviation of ten points.

The physical component of health related quality of life is central in patients with RA. In addition to physical function, the broader PCS

measure also captures, for example, limitations in role and social activities.

When we compare the activities assessed between the HAQ and the SF-36 physical function domain, this slide provides an example of some similarities and some differences. The HAQ asks about the performance of activities of daily living and instrumental activities, such as getting in and out of a car and reaching overhead.

On the other hand, the SF-36 asks about discretionary activities, such as walking greater than a mile or climbing several sets of steps; activities that would be important to patients who had little impairment in physical function. In other words, the HAQ asks greater detail of physical function, whereas the SF-36 asks broader or higher level questions of physical function.

When we look at the relationship between HAQ and SF-36, data from longitudinal studies and recent randomized clinical trials of new DMARDS demonstrate a high correlation between improvements in physical function as measured by the HAQ and

health related quality of life as measured by the physical function domain and PCS score of the SF-36.

These coefficients demonstrate that improvement in physical function closely correlates with improvement in health related quality of life.

Now I'd like to move and provide a brief review of some terminology that will be used throughout the presentation today. When examining mean changes across treatment groups, it is important to understand what these may mean to an individual patient.

The minimum clinically important difference, or MCID, indicates the amount of improvement that is perceptible to an individual patient and considered clinically meaningful. Although the MCID is relevant on an individual patient basis, when group median and mean scores well exceed MCID, it can be estimated that a majority of the treatment group will attain clinically important improvements.

This table summarizes the MCID values that we use for the HAQ DI, PET Top 5, and SF-36

based on statistical analyses of recent randomized controlled trials. These improvements are a negative .22 for the HAQ DI, a negative five points for the PET Top 5, a positive five to ten points for the SF-36 domains, and a positive 2.5 to five points for the PCS and MCS of the SF-36.

The second term that will be used throughout the presentation today is the number needed to treat, or NNT. The NNT is the number of patients required to receive a treatment with the agent in question to obtain one additional benefit beyond that achieved with the comparator or standard therapy.

Individual patient responses for HAQ, SF-36, and PET Top 5 can be distributed based on MCID values by treatment group. Proportions are calculated yielding a net benefit. The NNT is then expressed as the reciprocal of the net benefit.

The NNT approach is a practical and attractive way to express randomized controlled trial results as it informs the physician how much must be expended to achieve a desired benefit.

In closing, in a chronic and debilitating disease, such as rheumatoid arthritis, ameliorating the signs and symptoms is a major treatment goal. However, another very important and meaningful goal to the patient is improving and maintaining their physical function and health related quality of life.

Now, I'd like to introduce Dr. Karen Simpson who will present the leflunomide physical function and health related quality of life efficacy data.

DR. SIMPSON: Thank you, Mr. Doyle.

I will be reviewing the physical function and health related quality of life data from the three Phase 3 pivotal studies of leflunomide.

First, I'd like to provide some orientation to the studies and the patient populations.

The Phase 3 leflunomide pivotal studies included the 24 month US301 protocol, the multinational MN301 protocol with its series of two

extension studies called MN303 and MN305, totaling 24 months of blinded treatment, and the MN302 multinational protocol with its extension, also totaling 24 months of double blinded treatment.

US301 was a placebo controlled trial designed to show superiority of leflunomide to placebo and to compare leflunomide to methotrexate at the primary 12 month endpoint. US301 predetermined that placebo would not be analyzed beyond the 12 month primary endpoint due to the expected high number of placebo dropouts.

MN301 was a placebo controlled trial designed to show superior of leflunomide to placebo and to compare leflunomide and sulfasalazine as six months. All placebo patients were offered active treatment at six months, at which time placebo was switched in blinded fashion to sulfasalazine.

The placebo switched patients were thereafter excluded from subsequent analysis.

MN302 was not a placebo controlled trial. It was an active controlled comparison of leflunomide and methotrexate at 12 months.

Throughout the presentation I will be referring to the intent to treat cohort or ITT cohort and to the year two cohort of the studies depicted here graphically.

The ITT cohort for each study is the population of patients who were randomized and received a dose of study medication. The ITT cohort was analyzed at the primary analysis endpoint for each study designated by the bolded lines.

This was done to demonstrate the efficacy of leflunomide at six months in one study and at 12 months in two additional studies. The leflunomide ITT cohorts of these studies totaled 824 patients.

The year two cohort is the subset of patients who continued for a second year of therapy either by continuing in the 24 month US301 protocol or by enrolling in the second year extension studies in Europe.

Patients were not required to be responders in order to be in the year two cohort. The year two cohort is used to evaluate the

maintenance of effect, and the year two cohorts of these studies totaled 450 patients.

The statistical analysis plans for these protocols provided that the ACR 20 response was the primary efficacy measure in all three protocols. This is the standard efficacy measure used by the FDA to determine efficacy in rheumatoid arthritis clinical trials.

The ACR 20 responder rate was analyzed at the primary endpoint of each study, six months in MN301, 12 months in US301, and 12 months in MN302.

Secondary outcomes were X-ray and physical function. The primary endpoints for X-ray and physical function analyses were at the same primary analyses endpoints used for the ACR 20. All studies had ACR response, X-ray and physical function data at six, 12, and 24 months. US301 expanded the physical function evaluation to include health related quality of life.

The ACR response and X-ray data from the six and 12 month analyses of the intent to treat populations for these studies formed the basis for

leflunomide's indications to reduce signs and symptoms and retard structural damage in rheumatoid arthritis patients.

Today the physical function data will first be presented for the ITT cohort demonstrating the benefits at the primary analysis endpoint for each study, six months for MN303, 12 months for US301 and MN302.

Analyses will then be presented for the year two cohort. The year two cohort analysis was designed to determine if the benefits evident at 12 months were maintained in patients continuing a second year of active double blinded treatment.

The analyses are intent to treat using last observation carried forward and are performed in those patients with a baseline value and an endpoint or exit value for the efficacy measure being evaluated.

US301, the placebo controlled comparison of leflunomide and methotrexate, enrolled 508 patients. Methotrexate dose was 7.5 milligrams to 15 milligrams in the first year, with an increase

allowed to 20 milligrams per year in year two. The median dose was 15 milligrams per week in both years.

Ninety-eight percent of the patients received folate supplementation due to the blinded methotrexate treatment arm.

MN301, the placebo controlled comparison of sulfasalazine and leflunomide, enrolled 358 patients. Sulfasalazine maintenance dose was two grams per day after escalation from an initial starting dose of .5 grams per day.

The MN301 study and its extensions were conducted primarily in Europe, but also in South Africa and Australia.

MN302 was designed to show equivalence between leflunomide and methotrexate at 12 months with a sample size estimated to be 750 patients. Nine hundred ninety-nine patients were actually enrolled.

Methotrexate dose was 7.5 milligrams per week, with increase to 15 milligrams per week at the discretion of the investigator.

Comparing doses of methotrexate between the MN302 study and the US301 study, we can see that the median methotrexate dose was higher in the US301 study in which 98 percent of the patients received folate compared to only ten percent of patients in MN302 trial usually initiated after an adverse event had occurred.

All of the studies required that patients have active rheumatoid arthritis and be naive to the active comparator. Entry criteria did not limit the population to any particular maximum disease duration.

Disease characteristics and disposition were somewhat different among the protocol populations, and I will now review these.

In the US301 study, completion rates at 12 months and 24 months were similar in the leflunomide and methotrexate treatment groups. The 98 leflunomide and 101 methotrexate patients and 36 placebo patients who completed 12 months continued into the second year of treatment in this two year study.

These are called the year two cohort, and I've abbreviated it here as Y2C.

As expected, few placebo entered a second year of treatment. A high percentage of the year two cohorts, 85 percent for leflunomide and 79 percent for methotrexate completed 24 months of treatment.

Of the patients who withdrew in the first year, those who withdrew at or after four months, who had documented lack of efficacy, were allowed to enter a separate 12 month alternate therapy phase of the protocol not included in the analysis.

In terms of overall protocol completion, 52 percent and 51 percent of the active treatment patients and 48 percent of placebo patients either completed the 24 month study or completed 12 months of alternate therapy.

The effect of having an alternate therapy phase available for patients to enter can be reflected in this curve of discontinuations over time due to lack of efficacy in the ITT cohort of

the US301 study. It is clear that most of the patients exiting for lack of efficacy did so at or after four months when they could enter the alternate therapy phase.

In MN301, 72 percent of leflunomide and 62 percent of sulfasalazine patients completed the six month study. There was no placebo treatment being six months at which time placebo patients were switched to sulfasalazine as I've previously described, and they were thereafter not included in this analysis.

Completion rates at 12 months in the MN303 extension and at 24 months in the further MN305 extension were similar between leflunomide and sulfasalazine.

The 60 patients in each treatment group who enrolled in the second year extension study, MN305, abbreviated here at Y2C, comprises the year two cohort, and of those patients, a high proportion, 88 percent for leflunomide and 78 percent for sulfasalazine completed the 24 months.

Completion rates at 12 months in MN302

and 24 months in the MN304 extension were higher than in the other studies, as might be expected in an active controlled trial such as this where placebo treatment was not an issue.

The 292 leflunomide and 320 methotrexate patients who enrolled in the second year MN305 extension study comprised the year two cohort, and again, as in the other studies, a high proportion, 88 percent for leflunomide and 87 percent for methotrexate completed the 24 months.

Baseline characteristics show some differences among the ITT populations of the studies. In MN302, more patients had a shorter disease duration, up to two years, and fewer patients had a long disease duration of greater than ten years.

This is reflected in the much lower mean disease duration in the MN302 population despite a higher number of mean DMARDs in the past and a lower number not on previous DMARDs.

Taken together, these features suggest overall more aggressive disease in the MN302

population.

Baseline HAQ disability index scores show the most impairment in function in the MN301 population, as might be expected with their longer disease duration. In the MN302 population, the baseline HAQ disability index was already similar to the MN301 baseline disability index even though the disease duration was much shorter, another suggestion of more aggressive disease in the MN302 population.

Baseline demographics and disease characteristics for the year two cohorts from these three protocols were similar to the intent to treat populations. So these baseline features did not distinguish the patients continuing for a second year of treatment from those in the initial ITT population.

Now that I have described the studies and the populations, I will review the results for patient reported outcomes of physical function and health related quality of life. In order to evaluate the effect of leflunomide on physical

function, it was first necessary to demonstrate the efficacy with regard to overall signs and symptoms. leflunomide has been demonstrated to reduce signs and symptoms of rheumatoid arthritis as indicated in the product labeling. The graphic shows the time course of the ACR 20 responder rate by last observation carried forward to the 12 month primary endpoint of US301.

US301 was a 24 month protocol, and therefore, it's appropriate to extend the ITT analysis out to 24 months, demonstrating the benefit evident at 12 months was sustained in a second year of blinded active treatment.

As prespecified in the protocol, placebo data were not included in the analysis after 12 months due to the expected low numbers of placebo patients remaining in the study.

Now, I will review the patient reported outcomes of physical function and health related quality of life in the two placebo and active controlled trials and the one active controlled trial that I have just described.

For each outcome measure, HAQ, or SF-36, the ITT cohort data will be presented first in order to demonstrate improvement with leflunomide treatment at the six or 12 month primary endpoint for each study. This will be followed by the year two cohort analysis at 12 and 24 months in order to demonstrate that the benefit evident at 12 months was sustained in patients continuing a second year of blinded active treatment.

The HAQ instrument was accompanied by a visual analogue scale to allow the patient to indicate which activities were most important to them and which were most difficult for them, and these data were used to analyze the PET, or problem elicitation technique, scores.

In addition, the shorter, simpler, modified version of the HAQ, called the modified HAQ mentioned by Dr. Fries, was done on a monthly basis at each visit and was used to calculate the ACR 20 responder rate.

The HAQ disability index was done at months six, 12, and 24 in all of the studies, and it

is the HAQ disability index, our primary measure of physical function, that I'm presenting.

This graphic will show the mean change in the HAQ disability index in the ITT populations at the six or 12 month primary endpoints across all three Phase 3 studies. Improvement is a negative change from baseline. The numbers in parentheses represent the patients with a valid HAQ questionnaire at baseline and at the endpoint or early exit according to standard HAQ analysis procedures.

In US301, the improvement at 12 months is minus .45, and this was highly significant compared to placebo, which shows little change from baseline. The dotted line at .22 represents the minimum clinically important difference.

Improvement in the leflunomide treatment group exceeded the minimum clinically important difference by twofold.

The pattern is similar in the MN301 six month endpoint. Mean improvement in the leflunomide group is minus .56 and statistically significant

compared to placebo, again, little change being seen with placebo treatment. Both active treatments exceeded MCID.

In MN302, there was no placebo control. However, both leflunomide and methotrexate improved HAQ disability index from baseline. The improvement in the leflunomide treatment group was consistent with that observed in the other two studies. Mean changes in both active treatment groups well exceeded the MCID of .22.

In US301, because improvements in HAQ disability index were statistically significant at 12 months for both active treatments compared to placebo, we can compare changes in the individual HAQ subscales. Improvement with leflunomide treatment was statistically significant compared to placebo in all eight of the HAQ subscales.

These are the mean HAQ disability index scores over time in the leflunomide and methotrexate year two cohorts in US301. This pattern will be repeated in all three protocols, showing the improvement at six months and showing that the

improvement at 12 months was maintained at month 24.

These represent improvements in the leflunomide patients at 50 percent and the methotrexate patients of 31 percent. The percent of patients who achieved MCID is across the top, 71 percent for leflunomide and 59 percent for methotrexate.

To apply some perspective, an example of a patient with a baseline score of 1.2 might be a patient with some difficulty performing most daily activities and requiring, for instance, a jar opener to open jars or a bathroom bar to get on and off the toilet. Improving to a score of .6 might mean no difficulty performing most daily activities.

Similarly, in the MN301, 303, 305 series, the year two cohort patients showed maximum improvement at six months, which was sustained at 12 and 24 months. This represented a 46 percent improvement in the leflunomide year two cohort patients and a 37 percent improvement in the sulfasalazine year two cohort patients. Eighty percent and 71 percent of the year two cohorts

respectively achieved MCID.

The same pattern over time appears again in the MN302, 304 year two cohort showing the improvement in HAQ disability at six months and showing the improvement at 12 months to be maintained over 24 months.

The scores at 24 months represent 32 percent improvement in the leflunomide group and 37 percent improvement in the methotrexate group. Sixty-seven percent of the leflunomide and 73 percent of the methotrexate patients achieved MCID.

This graphic will show the same year two cohort, month 24, HAQ disability index data represented as mean change from baseline across the three studies. In US301, mean improvements in both treatment groups well exceeded the MCID. A similar pattern was observed again in MN301, in the MN305 extension study. With both leflunomide and sulfasalazine mean improvement from baseline well exceeded the MCID.

And in the MN302-304 year two cohorts, mean improvements from baseline in the leflunomide

and methotrexate treatment groups well exceeded the MCID.

To summarize the HAQ disability index data, the three studies demonstrated that leflunomide significantly improved physical function compared to placebo, in a placebo controlled six month trial, a placebo controlled 12 month trial, with further confirmation in a non-placebo controlled 12 month trial showing a consistent degree of improvement.

Improvement in physical function was maintained between month 12 and month 24 in patients continuing for a second year of leflunomide treatment.

The SF-36 generic measure broadens the definition of functional outcomes to reflect the impact of physical function on role and social participation and other important domains of health related quality of life. These domains were measured in the US301 study at baseline, month 12, and month 24, in addition to the HAQ instrument.

This graphic was previously shown by Mr.

Doyle, and I show it again to depict the baseline scores for each domain of the SF-36 for the entire U.S. 301 study population compared to the age and gender adjusted U.S. norms. Marked decrements in role physical, physical function, and bodily pain are evident compared with the U.S. norms.

So active rheumatoid arthritis affects all domains of the health related quality of life, although the physical domains reveal the most impact of the disease.

As you may recall, in the SF-36, a positive change indicates improvement. The dotted lines mark a change of five to ten points considered in the literature to represent a range of MCID. For the placebo group, mean changes from baseline in the intent to treat cohort at 12 months showed little or no improvement in most of the domains, with the exception of role physical.

Change scores reached or exceeded the MCID range in seven of the eight domains with leflunomide treatment and five of the eight domains with methotrexate treatment. Improvements with

leflunomide treatment were statistically greater than placebo in five of eight domains: physical function, bodily pain, general health perception, vitality, and social function.

This graphic will show the SF-36 domain scores at 24 months in relationship to the year two cohort baseline values and the U.S. norms simultaneously, providing another way to understand what the observed changes in domain scores might mean in terms of clinically meaningful improvement.

The white line indicates the baseline domain scores for the year two cohorts of both active treatment groups. The red line indicates the age and gender adjusted U.S. norms. The bars show SF-36 domain scores at 24 months, for the leflunomide year two cohort in blue and the methotrexate year two cohort in yellow. Domain scores in the leflunomide at 24 months approach or meet the U.S. norms in the eight domains of the health related quality of life.

Similarly, we can use the same type of representation to look at the leflunomide year two

cohort at month 12 and month 24. Month 12 is in the light bar, and month 24 is in the blue bar.

This shows that the improvements had already occurred at month 12 in each domain, and they were maintained at month 24.

The SF-36 domain data show that the improvement in physical function demonstrated by the HAQ disability index at six and 12 months and maintained over 24 months is reflected similarly in improvements in health related quality of life, not just in domains of physical function, role physical, and bodily pain, but also vitality, general health perception, social function, role emotional, and mental health.

The SF-36 physical component summary score, or PCS, for leflunomide and methotrexate year two cohorts are shown at baseline, month 12, and month 24. Baseline PCS scores 30.9 for leflunomide and 30.2 for methotrexate, are two standard deviations below the U.S. norm, and provide much room for improvement. It is evident that improvements at 12 months and 24 months in the year

two cohorts are remarkable, and in fact, PCS scores improve more than ten points, which is one standard deviation unit, and are within a standard deviation unit below the U.S. norm in the leflunomide treated patients.

For reference, the MCID for the PCS score in the literature is a change of 2.5 to five points.

The SF-36 data, like the SF-36 domain data, support the HAQ disability index data in demonstrating the improvement in physical function with leflunomide and the maintenance of benefit during a second year of treatment. The SF-36 results also demonstrate that the beneficial effect of improved physical function is substantial and reflected in health related quality of life.

This degree of improvement would potentially mean, for example, that a patient not able to work could possibly be able to return to work.

To look at the improvements in physical function and health related quality of life in a

different way, we can use definitions of MCID to calculate the number needed to treat to provide the defined benefit to one additional patient compared to placebo. The lower the NNT, the better.

NNT is provided here for the HAQ disability index and for the PCS score of the SF-36 for which a conservative MCID estimate of five was used. For both leflunomide and methotrexate, the NNTs are quite low for these measures.

Another way to examine patient reported changes in physical function and health related quality of life is to look at these changes in relation to the health transition question included in the SF-36 instrument. The health transition question asks: compared to one year ago, how would you rate your health in general now?

In those patients receiving leflunomide who achieved MCID in the HAQ disability index, 91 percent stated in the transition question that they had improved.

Conversely, of those who said in the transition question that they had improved, 75

percent had achieved MCID in the HAQ disability index. This pattern of agreement was very similar for the PCS score of SF-36.

Correlations between improvement in HAQ disability index and improvement in health related quality of life by SF-36 in longitudinal observational studies and recent randomized clinical trials was previously shown by Mr. Doyle. This plot shows the correlation between improvement in the HAQ disability index and improvement in the SF-36 physical component score in the US301 study in the leflunomide patients.

Another perspective on the physical function data is to look at the percentage of patients who have improvement or no change in HAQ disability index across the three studies. This is shown for the year two cohorts of the studies.

A very stringent definition used changed scores of less than or equal to zero to indicate no deterioration. It is evident that a high percentage of patients in all active treatment groups reported either improvement or no change in the ability to

perform physical activities.

In the leflunomide year two cohorts, 84 percent, 86 percent, and 74 percent of patients had improvement or no loss in physical function over two years of treatment.

The HAQ disability index and SF-36 physical component summary score in US301 side by side show that the proportion of patients with improvement or no change in physical function was similar for the HAQ disability index and the SF-36 PCS score. Eighty-four percent and 80 percent of the leflunomide patients who entered the second year of treatment had improvement or no loss in physical function over two years of treatment.

A number of conclusions can be drawn. Leflunomide is known to provide significant improvement in clinical signs and symptoms of rheumatoid arthritis and to retard structural joint damage, and these benefits are reflected in the product labeling.

But just as importantly, leflunomide improves physical function, and the benefit at 12

months is maintained in patients continuing a second year of treatment. The improved physical function is reflected also in improved health related quality of life and is clinically meaningful to patients.

The improved physical function was seen consistently across three Phase 3 studies with two year double blind data sets.

Thank you, and I will now return the podium to Dr. Michael Rozycki.

DR. ROZYCKI: Thank you, Dr. Simpson.

I would just like to wrap up with two slides to summarize what we've presented this morning with a number of summary bullets.

First of all, Aventis believes that improvement in physical function is the appropriate term for claims for physical function for the reasons discussed by Dr. Fries this morning earlier.

Aventis believes that 12 months of data is adequate to establish a claim for improvement in physical function. We see clinical improvement as early as six weeks after initiating treatment of leflunomide, and we see statistically significant

improvement at six or 12 months in the ITT cohort data, and benefits are maintained at 24 months in the vast majority of patients who continue on therapy.

Data indicate that placebo controlled trials are not necessarily appropriate for demonstration maintenance effect because of the dropout rate, and finally, results for patient reported outcome measures were consistent across the three studies involving a total of 824 patients, of whom 450 entered the second year of treatment.

In Study US301, which used multiple patient reported outcome measures, the HAQ and the SF-36, in particular, efficacy results were consistent across measures.

This concludes Aventis' efficacy presentation. We can accept questions now or will there be a break?

CHAIRPERSON ABRAMSON: Right. What we would is if members of the committee have specific questions for clarification of the speakers, we would take a few minutes to do that, and then we'll

have a discussion more openly subsequent to that .

Dr. Elashoff.

DR. ELASHOFF: I'd much rather ask them after a short break, but if we have to do it this way --

(Laughter.)

CHAIRPERSON ABRAMSON: Make the question short and then we'll take a short break.

(Laughter.)

DR. ELASHOFF: I have three questions. The first one is with respect to Study MN302. It was stated that the study was planned to have 700, but it ended up with 1,000 essentially. Why was that change made?

DR. ROZYCKI: I think probably Dr. Vibeke Strand is the best person to answer that question, and she'll take that question from the microphone on the other side of the room.

DR. STRAND: Very briefly, accrual was low, and so there was additional efforts to accrue more patients, and in fact, it was over subscribed. That led, of course, to there being statistically

significant differences between methotrexate and leflunomide, some of which would not be considered clinically meaningful. The ACR 20 criteria is statistically different, although the difference in the tender joint counts, for instance, were only three and in swollen joint counts only 1.8 between treatment groups, and that would explain, too, why the HAQ disability index differed by only ten points.

DR. ELASHOFF: My second question has to do with Slide MM61. It appears from the way they are labeled that the three different studies were originally on different scales, and what they were put on here, it was done as if they were on the same scale, but they are not. So that's a misleading slide.

DR. ROZYCKI: I think, Dr. Simpson, do you want to?

DR. ELASHOFF: Because the .6 and .56 are much further apart than the .48 and the .56. So there's just something wrong with the scale on that, but I just want to point that out. I don't need an

answer for that.

DR. ROZYCKI: Okay.

DR. ELASHOFF: The next thing has to do with the business of last observation carried forward. If HAQ was only done at six, 12 and 24 months, what last observation was carried forward if somebody left at three months or if somebody left at five months or at seven months, for example?

DR. ROZYCKI: Dr. Strand will answer this questions as well.

DR. STRAND: As Dr. Simpson mentioned, a modified HAQ was used in the U.S. study every month, and it was used to calculate ACR criteria, and the HAQ was administered in the MN studies every month, and the mean HAQ score was used to calculate the ACR criteria.

The full HAQ disability index was scored at zero, six, 12, 18, and 24 months to look at this maintenance of benefit in the year two cohorts and also look at the effect on physical function in the ITT. So last value carried forward would be zero to six months, from six to 12 months, from 12 to 18

months, and from 18 to 24 months.

But the year two cohorts were defined as patients who entered the second year of treatment. So the most that their ITT analysis would be carried forward would be a full 12 months to 24 months, and as you may have seen already, approximately 85 percent of the year two cohorts in the leflunomide treatment groups completed the second 12 months of treatment.

DR. ELASHOFF: Dr. Simpson said something about people who left early might have had an exit HAQ. Is that not true? You didn't mention that.

Could we actually have some sort of slide that makes this really clear for each study exactly when the HAQs were done and when they weren't?

DR. DAY: My question is related to that, if I could. There are so many multiple measures and they're taken at many points in time, which is good, but could somebody summarize for us in a given study how many different times an

individual patient was tested? Because you can have patient expectation with multiple uses of these instruments and so on.

So for a study with the maximum amount of testing with the maximum number of instruments, how many times were patients tested?

DR. ROZYCKI: Dr. Simpson.

CHAIRPERSON ABRAMSON: Before you -- I'm sorry. Obviously Dr. Elashoff was right. The complexity of the questions and the need to get into some depth with these particular issues, I think, will warrant the discussion time. So rather than do as I first intended, which was to get some crisp clarifications, what we'll do is we will hold that question and we can get a clarification on the slide that Dr. Elashoff had commented upon, and when we get to the discussion of the questions, the committee members will have a chance to get into real depth where I think we're heading with these kinds of questions.

So we will take a break now for ten minutes, come back at no later than a quarter to 11

with the presentation by Dr. Choi.

Thank you.

(Whereupon, the foregoing matter went off the record at 10;33 a.m. and went back on the record at 10:49 a.m.)

CHAIRPERSON ABRAMSON: We're about to resume, and we're waiting for all of the committee members to return.

All right. What we plan to do before Dr. Choi's presentation is to ask Aventis to simply respond to the last question that was on the floor, and after we get a clarification of that, we'll have Dr. Choi's presentation and then discuss the questions.

And there will be ample time for information to be obtained from the sponsor as needed to inform the discussion of these questions.

So I'd like to call on Dr. Strand to respond to the last question that was on the table before the break.

DR. STRAND: We just wanted to quickly respond for clarification only. Dr. Elashoff was

correct. We do have the numbers at the bottom of the bars so that people should know what the actual numbers are, but this slide has been corrected, and we apologize for the error.

And for the next point of clarification only, we wanted to point out that this is when the tests are performed in all of the studies. It's a standard design in randomized controlled trials in rheumatoid arthritis.

Of course, there's an endpoint determination. So, in fact, all of these values are last value carried forward to the endpoint or study exit, and study exit then would be carried forward.

And Dr. John Ware, who is with us today, would like to discuss at a later time point this business of multiple testing in terms of patient reported outcomes, but not at this time.

Thank you.

CHAIRPERSON ABRAMSON: Thank you, Dr. Strand.

We will now go back to the agenda and ask Dr. Choi to present on the statistics relevant

to these discussions.

DR. CHOI: Good morning. I'm Suktae Choi, a statistician in FDA.

This is title of my presentation. I change title. "Statistical Issues in the Analysis of Two Year HAQ for Arava."

This presentation will be about the problems of statistical analysis for duration of two year clinical studies due to high rate of early dropouts. It will be based on the real examples which are two years studies in Arava performed by Aventis.

Aventis submitted three studies with a duration of two years, one U.S. and two European studies. The U.S. study with the protocol number of US301 had three treatment groups, leflunomide, placebo, and methotrexate. It was a randomized, parallel, double blind study followed for two years.

One of the special features of this study was that non-responder subjects were switched on treatment at week 16. Non-responder in leflunomide group had to switch to methotrexate, and

non-responders in placebo and methotrexate group had to switch to leflunomide.

In the efficacy analysis, the three switch patients were considered as dropout at week 16.

The two European studies were very similar to U.S. study, except the treatment group. MN301, 303, 305 used a sulfasalazine as an active comparator and placebo treated groups switched their treatment to sulfasalazine at weeks 24 and excluded from two year analysis, and MN302-304 used methotrexate as active comparator, and there was no placebo treated group.

This presentation will be focused on the U.S. study because these studies provide similar issues and similar conclusions in efficacy.

The efficacy endpoint reviewed for year two or HAQ and MHAQ changed from baseline at the end of year two. Therefore, the proportion of alterations at the end of year two is very important.

For statistical analysis, the analysis

of covariates was used with LOCF method for imputation for missing data.

This table shows the number of percentage of subjects who were randomized and who completed two year duration. Overall 508 subjects were randomized and 190 were for leflunomide; 128 were for placebo; and 190 were for methotrexate. It was three to two sampling as planned in the protocol.

At the end of year two only 190 subjects completed out of 508, which is only 37 percent. However, not every completers had HAQ measurements at the end of year two, but only 136 subjects had HAQ measurements at the end of year two, which is only 28 percent of 508 randomized subjects.

Therefore, when LOCF method was used, 20 percent of the data were observed at the end of year two and 72 percent of data were carried forward from previous measurements.

For the leflunomide treatment group, 32 percent of subjects had HAQ measurements at the end of year two, and for placebo only 17 of them had HAQ

measurement at the end of year two.

This chart shows the change from baseline in HAQ at two years. The solo circle represent mean of leflunomide treated group, and the vertical bar is the plus-minus one standard error. The white color is for placebo.

When the missing data were imputed by LOCF, leflunomide shows significantly better than placebo with very small p value. However, this LOCF data are a combination of two different types. One is completers who have HAQ measurements at the end of year two and others is carried forward from previous measurements.

If we analyze the data by these two types, it will be like this. The pair in the center are for completers for HAQ, which means the patients who had HAQ measurements at two years. Remember that this analysis is based on 28 percent of ITT who completed and have HAQ measurements at the end of year two.

The pair on the right side, the pair on the right side is for the imputed cases. That means

the subjects who did not have HAQ at year two. So they're carried forward from previous measurements using LOCF. Remember that this analysis is based on 72 percent of the ITT. Therefore, we can say that LOCF analysis result is determined by imputed cases more than completers.

As we see, these two results are very different. This implies that imputed data are possibly biased. This orange is for methotrexate, and as we see, this group is not consistent either.

Okay. Now we want to show where this imputed data are carried forward from. The concentration of patients remaining at each time point for HAQ, that means -- okay, for HAQ. The black solid circle is the line for leflunomide, and the white is for placebo. The orange is for methotrexate.

There are two big drops during the first year. The first one is at week 16, and it is a surprise because non-responders were switched in treatment at this time point. So many of them were excluded from the study. Especially the placebo

treated group shows a big drop.

The second big drop is at the end of first year, which is at 52 weeks. So we can see that among the drop-off subjects, most of the last HAQ measurements were from first year period. In other words, in the LOCF analysis imputed data, which is majority of ITT, are carried forward from first time, some time in first year period.

These are the reasons that the patient drop off from the study: lack of efficacy, adverse events, and voluntary withdrawal, and so on.

This chart shows HAQ scores change from baseline using LOCF for missing data. The black is leflunomide; white is placebo; the orange in methotrexate.

The HAQ was measured at six month, one year, and end of two years, and when they exit.

This is the same chart, but only with observed cases. In other words, LOCF was not applied so that missing data were not imputed.

As you see, these are very different, especially at the end of year two.

This chart shows MHAQ scores changed from baseline using LOCF, and this is the same graph but only with observed cases. For MHAQ these two graphs are more different than HAQ.

This time point is at week 16, right before too many subjects were excluded, dropped from the -- switched from the analysis, dropped from the analysis. As you see, these two graphs are not much different up to week 16.

In other words, week 16 is the latest time point that can provide the most robust analysis results.

In U.S. study, because of the high rate of dropouts, the validity of two year analysis with LOCF is problematic, and we can find the same problems in European studies.

This is the number of patients at year two for one of the European studies. As you see, the dropout rate is still high, and this is for the other European study. The dropout rate seems better than two other studies, but not enough to be valid.

So this is my conclusion in this

presentation. There are less than 30 percent of patients with measurement of year two HAQ. So high rate of missing data validity of two year analysis with imputation of year one data becomes problematic.

And this is the end of my presentation. Thank you.

CHAIRPERSON ABRAMSON: Thank you very much.

Are there questions from the committee for Dr. Choi?

(No response.)

DR. CHOI: Thank you.

CHAIRPERSON ABRAMSON: We will now move into addressing the questions framed for the committee, and the procedure will be that the committee will address segments of the questions, and when our discussion either needs to be informed by either the FDA or the sponsor, we will ask specific questions of either and ask for more information.

Let me begin by reading the questions

that were distributed. The "Guidance for Industry Clinical Development Programs for Drugs, Devices, and Biological Products for the Treatment of" RA, released in February 1999, includes the recommendations for the claim "prevention of disability." As noted in this guidance, studies should be two to five years in duration to support this claim.

Recent studies attempting to assess efficacy and durability based on placebo controlled or add-on therapy studies have identified limitations for proper conduct and interpretation of these studies because of high withdrawal rates. Therefore, FDA is considering a revision of this claim.

The health assessment questionnaire, HAQ, has been evaluated in a variety of clinical trials and settings over the years, particularly for physical function in activities of daily living. It is recognized in the RA guidance document as an adequately validated measure for use as the primary outcome measure in trials of physical function in

rheumatoid arthritis.

Question No. 1: In light of the available literature on the HAQ instrument, does the term "physical function" or "disability" better capture the clinically relevant information ascertained in this instrument?

And I think before the committee addresses that question specifically, Dr. Jeffrey Siegel -- I'd like Jeff to address the precedent in terms of the infliximab label, in terms of the use of "physical function" versus "disability."

MR. SIEGEL: Thank you very much.

I'm currently Acting Branch Chief in the Immunology and Infectious Diseases Branch, and I was reviewer for the Remicade improvement in physical function DOA supplement.

I just wanted to make a couple of points. First, the claim of prevention of disability in the guidance document was intended to do a number of things. One of them was to collect long-term data on new products for rheumatoid arthritis. We had thought when the guidance

document was initially formulated that what we would see in these long-term studies would be a worsening in the HAQ in untreated patients, and we hoped to see stabilization of the HAQ, a lack of progression of disability in treated patients.

It turns out as we've done clinical trials and measured HAQ, that's not what we've seen. The problem is that even in untreated patients over the time course of clinical trials, disability doesn't worse. The HAQ does not increase. It actually stays the same, and this has actually been well validated in a number of long-term studies, epidemiologic as well as clinical trial.

So when we have the first request to get a claim of improvement in physical function or prevention of disability from Centocor for Remicade, we found we couldn't look at that. We couldn't see prevention of an increase in HAQ.

Instead what we saw in the control group is there was a tendency to be flat, and then in the treated group, there was a decrease in the HAQ. So we thought that prevention of disability, a

prevention of this increase in HAQ that we expected to see was really not the basis of the data that we saw. Rather, it was an improvement in the HAQ. We thought that was better expressed as improvement in physical function.

So the way that we assessed this was to look at whether there was a clear reduction in the HAQ in the treated patients compared to placebo, and whether that improvement in HAQ was maintained after two years.

So I just wanted to mention that that was the basis for using the term "improvement in physical function" as opposed to "prevention of disability."

CHAIRPERSON ABRAMSON: Thank you very much.

I would ask members of the committee what their thoughts are on this term "physical function" versus disability. Dr. Williams.

DR. WILLIAMS: If the author of the HAQ prefers "physical function," I would support that.

(Laughter.)

CHAIRPERSON ABRAMSON: Awaiting Dr. Fries, do you want to?

DR. FRIES: I indicated great ambiguity and willingness to go along with the majority roll here.

(Laughter.)

CHAIRPERSON ABRAMSON: May I ask just for a clarification? You have described very eloquently the HAQ disability index and have shown data on that. How does one think about that term in the context of this question? Does that capture what we need to capture?

DR. FRIES: Well, I think that it does. I mean, just in terms of the continuity of what's been happening, I would probably prefer disability index and proscribe the use of the word "disability" unqualified so that you were talking HAQ DI or something. I think that gets you away from the blue parking sticker things and the payments and the on/off disability kind of thing. It allows you to say it's an index. It's a continuous variable, essentially a continuous variable, and so forth.

But I can make arguments for physical function or any of the other sorts of range of acceptable things that they're accentuating the positive. The disability index is accentuating the negative. So basically my preference would be call it the HAQ DI or something like that, but it's just a question of, I think, the precedent and so forth that has been set with other drugs. You want to be consistent across medications with regard to what your terminology is. So there are a lot of these considerations, I think.

I'll just parenthetically say in light of the last remarks, just to operationalize why the HAQ is flat, because it absolutely is, I mean, it goes up. If you saw our data earlier, our data showed that it goes up about .017 a year in stable populations, and the reason for that is that rheumatoid arthritis for clinicians here -- when it hits, you basically have a tendency to have some difficulty in everything.

Now, some difficulty in everything means you have a HAQ of one. So there's sort of this

instantaneous rise with early disease from zero, assuming the people were perfectly fine, to one, and thereafter then you have these random effects of the treatments which tend to balance each other out and maintain your numbers quite stably.

So I think the point was very well taken that you're looking for improvement and some kind of sustained improvement in individuals in terms of physical function or HAQ DI.

CHAIRPERSON ABRAMSON: So the current language is "improvement in physical function" in the label right now.

MR. SIEGEL: For Centocor.

CHAIRPERSON ABRAMSON: It's in the Centocor label. That's what I mean.

So that's the language that exists, and I guess a question for us to consider as a committee is is that the right phrase or should it be improvement in disability index or some other terminology.

Dr. Gibofsky.

DR. GIBOFSKY: I think we should keep

the term "HAQ disability index" for the instrument and say that it measures physical function. I'm much more comfortable in dealing with patients with rheumatoid arthritis in trying to help them assess their level of function than in trying to define their level of disability.

The connotation both clinically and from a patient perspective is quite different. So perhaps we can resolve the conundrum by keeping the term the "HAQ DI" for the instrument, but understand that it's measuring physical function.

CHAIRPERSON ABRAMSON: And the criteria though that someone needs to achieve a label of improvement in physical function is the HAQ DI, or I guess that's another missing piece in this discussion.

DR. GIBOFSKY: Well, that's the next question, yeah.

CHAIRPERSON ABRAMSON: Right.

DR. SIMON: Jim, could you comment? In this flatness of the HAQ response or measure, could part of it -- and some have suggested it might be --

related to an acquiescence to one's new life, meaning you get the disease, you deal with getting the disease, you become acquiescent to what's happened to you, and so thus the changes that then are measured after that are different because of the new world order that you're now sitting in.

Does that complacency to one's new life play a role with that measure?

DR. FRIES: No. The reason, I think maybe this is what John Ware wants to say, or maybe he wants to say something else, but we can tell. If people go off of medications, let's say you go off of your methotrexate. It just goes right back to where it was. I mean, you have a flare.

I mean, so it's clear that it isn't becoming immune to the questionnaire phenomenon because you see it go on. The next time you put the TNF alpha on, even though you've got HAQs going back the last 12 years, you still get, you know, the .4 to .6 drop with a new drug. You go off of it, and it goes back up.

So it's very sensitive in an ongoing

way, and part of that, I think, builds down to the way in which questions are constructed to be very specific. I indicated earlier bend down and pick up a piece of clothing from the floor. So your answer to that is very -- it's imbedded in the question, the function, the function is. And so you're not asking how you rate your health, very good, excellent, fair, poor, in which case you really can have some problems with it.

These are very specific tasks which tie to your ability to do actions.

CHAIRPERSON ABRAMSON: Dr. Manzi.

DR. MANZI: Jim, from somebody that doesn't use the HAQ and is not very familiar with it, how do you deal with attribution from other comorbidities?

So, for example, if somebody has an osteoporotic compression fracture, it may affect those things. How does that -- how do you interpret that?

DR. FRIES: Well, as I tried to indicate, you do that imperfectly. The only thing

that we use -- because one could ask the strictly generic thing and just treat anything else that happens as noise, you know, the congestive heart failure, the fractured hip or something like that, and we do try with the single word in the questionnaire to focus it on arthritis, recognizing that people will not always perfectly attribute that question.

But in general, the things that one might worry about with an instrument balance out with regard to change score measures because they're likely to occur systematically throughout. So I'd say that there's no perfection with any instrument, regardless of what it is or who's making those observations, but it's a really darn good instrument, as you see.

DR. BRANDT: Well, I think what Lee was getting at that Jim responded to was the difference between disability and handicap, and if a person never has to reach up for a five pound bag of sugar, that has no relevance, but that's inherent in all of this.

DR. FRIES: Well, there is a whole issue in terms of these instruments as to how your stem is set. The HAQ stem is are you able to. It's not do you, but it's are you able to, and it's an attempt to get around this exact point.

And, again, I would acknowledge lack of perfection, but the intent is to see if people who don't do something, and you try and put things that people do do or almost have to do in, but recognizing that the rest of them in kind of a virtual way are either able to or not able to.

CHAIRPERSON ABRAMSON: So just in terms of this 1(a), I think the sense is that disability is a complicated word with many connotations that we'd like to avoid and physical function is the word that we'd like to promote as you have, and I guess Question 1(b) begins to address how one defines that consistently across agents.

So let me read 1(b). Are the more recent derivatives, such as the modified health assessment questionnaire, MHAQ, and the multidimensional health assessment questionnaire,

MDHAQ, appropriate and validated endpoints and substitutes for the HAQ in this regard?

Who can we hear from? Who wants to comment on this?

Dr. Williams.

DR. WILLIAMS: Well, the HAQ is the most commonly used. I think that you can state that any validated disability index could be used. The emphasis should be on "validated."

And I'm not sure. Has MHAQ been validated now, Jim?

DR. FRIES: Yes. I would basically take Jim's point. Obviously I love the HAQ and have a self-interest in it in a sense, but I would not like to see a universe which was closed to innovation by sort of saying we have this or not.

I indicated that the MHAQ was less sensitive. There are parts of the MDHAQ that may be too sensitive. You know, I think it goes up to running two miles and things. You have to sort of fix the range, but I would think that people should look at ease and clarity of administration to all

populations and do their NNTs when you do your power calculations and consider the range of all of the validations, and if it's multi-ethnic, the availability of translations and different culturally adapted instruments and so forth, and then make your choice amongst instruments that met the criteria.

As you saw here, the SF-36 is designed for entirely different things. Nobody was thinking -- I know John can comment again -- but nobody was thinking about randomized controlled trials in rheumatology at that point, but it actually works better than number of tender joints.

So I mean, I think, again, it's the importance of moving toward what we're trying to do for patients that to me is more important than the specific instrument chosen.

CHAIRPERSON ABRAMSON: Other comments from the committee on this?

If I can use the Chairman's prerogative to ask two of the consultants who are really expert on this to make very brief comments on their

opinion, Dr. Strand and Dr. Wolfe and Dr. Hochberg. I just would like to hear very brief comments on these three instruments and your views of them apropos the question.

REAR ADM. KLEINMAN: Just to look at the data between modified HAQ and HAQ disability index from the US301 study, it showed very close correlations between the two, but the HAQ disability index is more sensitive to change, and we have published that, Tugwell, Bombardier and myself.

And then I will let Fred and Mark answer.

DR. FREDERICK WOLFE: We've actually published a paper comparing several instruments, and the measurement properties of the MHAQ and the HAQ are entirely different because of the way the questions were selected. The MHAQ and the HAQ in clinical trials work approximately equally at the level of disability that one sees in clinical trials, which is high.

But the MHAQ is a totally poor instrument when you get down to low levels of

functional disability. It has about 32 percent of people with rheumatoid arthritis will have a normal MHAQ score compared to about 12 with a HAQ and compared to almost none when one uses a very good score, which is the physical function score of the SF-36, and the SF-36 and the HAQ differ only at the extremes. They both perform just about as well.

As long as I'm up, I want to say one other thing about physical disability. I think that the main driver of the HAQ is pain, and if you were to remove pain, then the question of physical function, what's the residual physical function, is a different question.

See, I think HAQ measures -- so I would say that I think if you really want to measure physical function, you have physically measure it. But I would think that the term "function disability" which takes into consideration both pain and the physical aspects is correct, but the reason why the HAQ goes up and down so fast early in disease and with this is pain change, and pain, of course, drive physical function.

But if you mean permanent physical function or you mean transitory, then there's two different things.

CHAIRPERSON ABRAMSON: Mark, do you have something to add to the choices of HAQs?

(Laughter.)

DR. HOCHBERG: Well, I've had experience with both the HAQ, having worked in Aramis as an investigator, actually published on the HAQ in lupus because I know Dr. Manzi is a lupologist, and also used the MHAQ, although I don't have experience with the MDHAQ.

I can agree with some of what Fred said and the data that Dr. Strand just showed in that on average when administered to the same patient population, the mean scores for the MHAQ are lower than the mean scores for the HAQ, and consequently, you may see less change as was demonstrated in these data as well over time.

I think what Dr. Williams pointed out is that what you need is not only a valid instrument, but one which is reliable when administered and

responsive in a patient population.

I really don't have any more to say, but if the Chairman doesn't mind, I'd like to yield any additional time I might have spent to Dr. Ware.

(Laughter.)

CHAIRPERSON ABRAMSON: So moved. We'll hold Dr. Ware perhaps for later, but, no, I think we have enough input right now. You give an inch.

(Laughter.)

CHAIRPERSON ABRAMSON: So the comment is, I think, from the committee, and people can comment otherwise, that for clinical trials -- oh, I'm sorry, Dr. Anderson. I apologize.

DR. ANDERSON: Actually I would like to hear from Dr. Ware. Maybe it doesn't have to be right now, but I'm interested in, you know, what he might say about the use of the physical function scale or even the PCS in this context.

CHAIRPERSON ABRAMSON: Dr. Ware.

PARTICIPANT: Thank you, Dr. Anderson.

DR. AWARE: Thank you.

We've had two eloquent lectures already.

I won't give you another one, but cosmically speaking, we label tools what we want them to measure, and when we change our labels, it doesn't change either the content or the empirical validity of the tool, and we need to remember that.

The fact is the HAQ -- and it is a darn good instrument -- measures the same physical domain of health as does the PF domain scale in the SF-36. The two together measure about four of the six standard deviations that we now can measure with all physical functioning measures, including the other tools that Dr. Fries mentioned.

So the HAQ lowers into the worse states by about one standard deviation below the PF scale in the SF-36 domain, and the SF-36 relative to HAQ raises in the favorable direction about one standard deviation.

Together that's only four. We get from sports medicine even higher levels, and from FIM and other tools we get even lower levels.

With respect to the labeling, the labeling is very important, and it's a lot like

thermometers 200 years ago. I don't know how many of you know that the original Centigrade scale, water froze at 100 and boiled at zero, and it wasn't until after the death of Celsius that the physicists got all of the thermometers going in the same direction.

I think I prefer tools that are labeled in the direction of a high score. So if it's going to be a functioning measure, there's a lot to be said for scoring it, you know, positively.

But there empirically does not change, you know, with a linear transformation in one direction or another, but the important thing is that we standardize the content, as has already been said, and that we collect interpretation guidelines, and that we maintain comparability with the past. We don't want to cut ourselves off from all of the interpretation guidelines we have for these scales.

But the labels are very important, and I have a strong preference for the improvement in functioning because of all of the political issues worldwide. The world is moving away from disability

to participation in life as a more positive concept, and a lot can be said for talking about this physical domain as functioning.

Finally, what is the difference between the PCS? The PCS just adds additional layers to the onion. It goes beyond physical functioning as a domain, which is measured by HAQ very well and by PF, and into the implications of physical problems for social and role participation.

And you know, when we see differences as large as we see with this treatment, those implications are great, and they should be considered when we do the risk-benefit calculation.

Here's a slide, if it's helpful conceptually. I created this specifically after reading these clinical trial results. Basically the clinical outcome is the structural impairment which you understand better than I do.

The PF domain score and the HAQ DI score very much get at the implications of this for physical function, and what we get with the PCS is the rest of the health related component, the

physical component of health related quality of life, and it allows you or it confirms that the physical improvement in life is more than ambulation and walking.

You have a social life. You're much less likely to be limited at work or to be unable to work or to take more frequent rests at work. These are very large improvements, and I just think the physical component adds understanding to the implications for human life beyond the more specific physical domain effect that we see with the HAQ DI.

Thank you.

CHAIRPERSON ABRAMSON: Thank you very much, Dr. Ware.

So I guess with regard to Question 1(b) what I think we're hearing is that the HAQ seems to remain the gold standard and the most comprehensive among these, and I'm wondering if anyone on the committee would speak to some other issue or disagree with that in terms of the --

DR. WILLIAMS: Again, I would just restate that while I agree that the HAQ is most

commonly used, if they can show another validated disability index, it ought to be accepted as well.

DR. FRIES: And I hope I was making the same point. I mean, the question is it's a quest for excellence, and if we closed it off, we would basically be saying, well, you know, this is as good as it gets, and I don't think we can ever say that in any areas of scientific inquiry.

And so I really would argue along with Jim's thing that we would require validation, and then that validation would maybe not be totally specified, but it would clearly have to satisfy the FDA when the product came up for review. It would have to be defended that, in fact, it was a valid measure.

But I would tend to keep it open, and if in the review of the HAQ review, you'll see that we advocate coming down as much as possible to the HAQ DI an the SF-36 as standards to which you work and model from.

CHAIRPERSON ABRAMSON: Is there any other information that you'd like form the

committee?

DR. SIMON: Looks good to me.

CHAIRPERSON ABRAMSON: It's okay. Okay.

So we'll move on to the next page.

For this meeting, the committee has bene provided data evaluating the effects of leflunomide on physical function from clinical studies, including data at 12 and 24 month time points. The effects of patient withdrawals on last observation carried forward landmark analyses of an intent to treat population at these time points has been discussed.

The current guidance notes that studies should be two to five years in duration. The Advisory Committee deliberations in 1998 concluded that the controlled data at one year demonstrated improvement in physical function.

Similar one year controlled data, along with durability of response during the second year in those patients who responded at one year, have been used to support approval of one therapy for improvement in physical function, that is,

infliximab.

For the domain of disability or physical function, what duration of a superiority study, placebo or active comparator, is needed to robustly identify an improvement?

And before the committee addresses that, I'd like to ask Dr. Siegel one more time to just put this in the context of the prior label for infliximab.

Do you want to wait until the third point? Okay. I'm sorry.

All right. So for the domain of disability of physical function, what duration of a superiority study is needed to robustly identify an improvement?

Jim.

DR. WILLIAMS: I don't know that we have a solid answer, but I think that with the more effective treatments, particularly for rheumatoid arthritis, that the longer placebo stage is becoming less common, and I would say that if they can show a difference in four to six months and then show

durability of that change for a longer period of time, but not necessarily under a comparator, I would accept that.

DR. FRIES: I think it was left on. I'm sorry.

Yeah, I totally agree with that. I think that unless we have at least one example of a drug in which it is not sustained once it begins or we have a clinical feeling that all of a sudden we have some drug that we lose it with, then I think we really can infer a lot from the first six to 12 months.

I have a feeling that 12 months is going to be required for approval on a lot of things. So it may turn out to be the de facto standard. I would actually, like Jim, be happy or satisfied with something which was less than 12 months, but I don't think we have to go beyond 12 months.

DR. WILLIAMS: Less than 12 months, but show that it persists for perhaps 12 months, even though you're not under direct comparison with another agent.

CHAIRPERSON ABRAMSON: Dr. Elashoff.

DR. ELASHOFF: I would like to talk about the word "robustly" rather than any specific times because the last observation carried forward procedure for filling in missing data may be reasonable under certain assumptions about the response pattern and the dropout pattern, but it is extremely easy to show that it is biased in, for example, the situation where the placebo and the active drug might show the same pattern over time in the physical function, but for some other reason like pain or something else, the placebo group drops out earlier on the average.

Their last observation carried forward will look worse than the active drug even though if you were somehow able to keep them in, they were showing exactly the same pattern.

So the issue of interpreting data where so much of it, even in the shorter term, has been filled in is very problematic, and I think that needs to be addressed much more in depth even interpreting the first year data from these studies.

CHAIRPERSON ABRAMSON: Dr. Gibofsky.

DR. GIBOFSKY: I would share those concerns. I think even though, as we've heard, we might be moving towards acceptance of a standard of one year or even less, with the ability to show the improvement at one year, to the extent that that one year is achieved by filling in of holes with last observation carried forward, I think that would be problematic as Dr. Elashoff has indicated.

I'm rather struck by Dr. Choi's comment for the data that we looked at. Week 16 is the latest time point which produces the most robust benefits, and I would like someone to respond to that at this point.

CHAIRPERSON ABRAMSON: Any --

DR. SIMON: Any particular person?

(Laughter.)

CHAIRPERSON ABRAMSON: Perhaps Dr. Strand.

DR. STRAND: I would like to respond to it because I did design the study, and there's a misunderstanding here. First of all, non-responders

were not required to exit. Only if a patient asked to exit for lack of efficacy were they allowed to exit for documented lack of efficacy, which was the absence of an ACR 20 response, although the curves show that the majority of the placebo patients exited on or after 16 weeks, and at that 16 week time point, there were some additional exits over time.

I think there's some information here that's useful about this ITT LOCF, and I'm going to start with the year two because we've been talking about it, but I think Dr. Cook would like to point this out, too.

If I could have the slide up.

In fact, if you look at the people who drop out in placebo versus the people who stay in in placebo, they are a very different patient population, and it's actually statistically significant at 12 months that the people who stay in the study for 12 months -- that's 37 out of the original 118 -- were responders, and they were so despite having longer disease duration and having

failed more DMARDs.

And what you can see on this slide is that if you look a the month 24 completers, of which there are interestingly enough 21, they have the lowest baseline HAQ disability index, but they do have also the longest disease duration and about the same number of DMARDs failed.

If we go to the next slide, you can see that, in fact, the people who drop out are the ones who are actually deteriorating. The 55 percent actually have an increase in their HAQ disability index. So they are dropping out because they are not responding.

If they leave for safety, they show some improvement. If they leave for other reasons, they also show improvement, and the people who actually do stay in the study appear to be the placebo responders.

Now, this type of pattern is also seen in the active controls, but it basically does say that the imputation of the last value, while they're still in their initial treatment assignment is an

appropriate imputation, but, yes, the active and the placebo over time will start to approach each other, and in fact, the placebo responders start to look as if they have responses similar to methotrexate over 24 months in this particular study.

DR. GIBOFSKY: Do I take it you disagree with Dr. Choi's assertion about week 16 being the latest time point at which one sees the most robust results?

DR. STRAND: No. I'm simply saying that week 16, I would prefer to take it at six months.

CHAIRPERSON ABRAMSON: On this slide, Dr. Strand, there were 27 -- this is the 301.

DR. STRAND: Right.

CHAIRPERSON ABRAMSON: So there were 27 patients who completed the two years?

DR. STRAND: Believe it or not there were 27 who completed two years, and there were 14 who completed three years of blinded treatment in the extension protocol on placebo, and they were responders with improvement in X-ray and improvement in physical functions.

CHAIRPERSON ABRAMSON: But just the numbers. There were 190 patients entered at time zero for --

DR. STRAND: One hundred eighteen in the original placebo group; 128 when we added the Canadian patients.

CHAIRPERSON ABRAMSON: Okay.

DR. STRAND: Thirty-seven completed the first year, and 27 completed two years, and 14 completed three years.

CHAIRPERSON ABRAMSON: Dr. Williams.

DR. WILLIAMS: I think this illustrates the point I was trying to make, that if you have a difference in a placebo controlled trial, this place was four months. You may want to pick six months, but then after that you don't have to worry about carrying values forward.

Did that response continue at that level for a year? And it's not compared to anything else, and I think that would eliminate the problem of whether you eliminated all of your severely ill patients and, therefore, your last value carried

forward is not adequate or accurate.

But I think that really illustrates that at the end of a controlled period, we had a difference. That difference was maintained over the next two years. Whether it was maintained compared to placebo is statistically difficult to determine.

CHAIRPERSON ABRAMSON: Dr. Makuch.

DR. MAKUCH: Yeah, I had my light on. I was just still thinking.

I think the comment is that there really is -- and I don't know what the answer is -- there really is a tradeoff between trying to get the best estimate of the effect versus on the other side what you have then are patients dropping out over time, and then you're getting increased variability and noise and sort of a mixed signal.

So, I mean, I agree maybe perhaps a bit with Dr. Choi. I think 16 weeks is the purest estimate that one can get.

However, I think it's probably not a long enough time, and certainly I've been hearing, and I would concur that somewhere between six months

and one year is probably the idea time, and where that precise cutoff is is a bit difficult because it really is a tradeoff with the loss to follow-up.

If there aren't many losses to follow-up, I would then recommend highly the 12 month. If it is confounding though the issue, then I would back down towards the six month, but again, exactly where that is, I think, is difficult for us to say, and I would certainly put it out as just an interesting question for others to resolve with those points in mind.

CHAIRPERSON ABRAMSON: May I just ask a follow-up to that?

The dilemma perhaps is that we have two issues. Is a 16 week time point a relevant outcome time point?

And then at two years, what is an appropriate number of people that need to be followed to complete two years versus the LOCF?

And so we have according to Dr. Choi only 28 percent of the people who completed 16 weeks being actually observed through two years, and I

guess I just would like to get a sense of the committee what that number means to them and what is a reasonable expectation to evaluate.

Dr. Makuch and then Dr. Elashoff.

DR. MAKUCH: It is interesting, and I guess I'm just going to make a generic remark. The generic remark is actually I think that what Dr. Choi did and what the Aventis people did is somewhat different in the sense that they are looking at the data from -- primarily looking at the maintenance issue, even though they did look at the six or 12 month data as well.

But I think looking at the maintenance issue. Given that you were doing well at one year, is that maintained over time? Very different than what Dr. Troy was doing where he was looking at from baseline going forward.

And so it's a very different, yet subtle distinction where he's saying is there a difference between the two groups from the get-go over a two year period as opposed to, I think, looking at conditionally at one year these are the data that we

have. Is it maintained?

Two different questions, and I really think that both analyses address it in a probably correct to some level in addressing those two somewhat distinct issues.

So I think both of the analyses are valid. I think Dr. Choi to me presented interesting analyses. Again, the further out you go from baseline, if you're looking at this overall effect from baseline, that the further out you go, the more problematic the results become, and that would sort of be my overall interpretation of what he was suggesting, and again, the precise time point then for looking at overall differences really then I think is a function of how much you're willing to go out before the loss to follow-up starts just deteriorating your results too much.

DR. ELASHOFF: Even starting at the one year period and using the one year follow-up from there to two years, they were using last observation carried forward, and in that case, it will make things appear to be stable even if they perhaps

weren't because if the person leaves the trial when they're not looking stable anymore and you're still using the last observation carried forward.

And in regard to that, I wanted to remind people about the slide that Dr. Fries put up, which suggests that things may turn around at some other time point. So we need to be using an analysis which will allow us to see if that's happening.

And last observation carried forward will tend to obscure that.

CHAIRPERSON ABRAMSON: Dr. Anderson.

DR. ANDERSON: Yes. I would like to see some other analyses. I know there are quite a lot of them there, but some analyses that were sensitivity analyses, and I would have more confidence in the results if we saw those.

In particular, people who dropped out at 16 weeks, they didn't really drop out. Many of them had a treatment change, and if there was an analysis by group that they were originally randomized to, regardless of what happened later on, and then used

the, you know, actual, not last observation carried forward HAQ scores that may have sometimes been obtained on a different treatment, that would be interesting to see, and that would be one way of assessing the strength of the results.

And there are sensitivity analyses, too, that can be done under different assumptions about what happens to HAQ, say, for people who drop out for different reasons.

So those sorts of analyses, I think, might serve to bolster the case.

MS. McBRAIR: Just in relationship to the time of placebo, I would just encourage people to keep it to a minimum. While patients are glad to advance science, they are possibly unable to function, living in severe pain, losing jobs, having impact on their families, having permanent joint damage occur.

So whatever the scientists deem as appropriate and scientifically okay would be okay with us, but I think there are other comparators now and other choices that people can use that I would

just encourage their group to consider.

DR. MAKUCH: One other comment. There are a lot of very bright biostatisticians in this room, and I think that the design of the studies in terms of when you stop the placebo and then cross them over to active treatment does not necessarily have to then affect the analyses.

There are other analyses in which one can make, and I guess this is follow up on Dr. Anderson, that you can make use of all the patients in the study with the variable follow-up, and that there are more complex methods available that then can do that. They should not be linked necessarily though to the actual treatment period for placebo, and that nevertheless you can then have a longer time at which the analysis is based in terms of the endpoint analysis without having the patients themselves to necessarily have to go through a long period of receiving placebo.

So there are ways, I think, to look at this question, and again, I guess there are additional analyses. I wouldn't want to see any

more today, but there are additional analyses that I think one could do that would really make use of the data in a more full way.

CHAIRPERSON ABRAMSON: Just to pick up on that and come back to the specific question in the context of rheumatoid arthritis, what duration of a randomized trial would be necessary to be sure that you've had the possibility to observe a sustained effect? And we've seen some 16 week data, and I'm just curious what the committee members think about what -- and perhaps I'll direct it specifically to Jim, Dr. Fries.

Using the HAQ disability index, what is the minimum number of months that you need to have a randomized trial to know that you've had an effect that is sustainable and real?

DR. FRIES: I think you have to go to the natural history of disease as shown by the observational trials, which is really why I was trying to show you at the 84 month data, and that there is some period of time.

But that data, and as far as I know,

there's no exception to it or not contrary data, would suggest that you can actually establish it quite early, as Jim is suggesting, and that it will be then continued for at least the periods of time that we're talking about.

If we had the additional thought that two years was a good time and now we find there are practical difficulties in going two years, the idea that you could predict in six months the two year data, I think, is a very strong suggestion from the other data.

So I'm really very close to where Jim Williams is on this, saying that it would be nice to just kind of set that point, whatever it is. Maybe it's a six month thing; to get a little bit farther than the 16 weeks, and then you just track it in those patients to see if there's regression or what we call reprogression.

CHAIRPERSON ABRAMSON: Dr. Williams.

DR. WILLIAMS: A lot of that depends on how rapid the drug works. If you have a treatment that works within a couple of weeks, you're going to

be able to identify it early, but if you're looking at gold, it may take you several months before you're going to see it.

So I think my own personal preference would be six months, but I don't have any real foundation for that, except that that would probably pick up the slowest one, which is gold.

DR. STRAND: Well, I would like to clarify. If you would like to see, we can show you the baseline characteristics and the HAQ responses of the early dropouts for the active treatment groups in the US301 study, which I think will illustrate a similar kind of a pattern that I showed you with placebo.

I will remind you that the patients who chose to enter the extension step protocols in Europe were about evenly divided between lack of efficacy, safety, and other reasons.

And then in data that we haven't shown because there's no time to, of course, even in these enriched cohorts for responses in the year two, these patients have ACR 20s of 70 percent to 77

percent, not 100 percent.

In other words, patients are staying in these protocols even if they're not ACR 20 responders. So there's a variety of reasons why either they're staying in the study or they're leaving the study, which doesn't necessarily reflect entirely either lack of efficacy or safety.

So I think that that's a point. Now, we did not feel it was appropriate to impute data over 24 months.

CHAIRPERSON ABRAMSON: May I just pause for a second? We need to come back to that later on in the question.

DR. STRAND: Okay.

CHAIRPERSON ABRAMSON: Because I think that's going to be a very important issue to really understand the data, but maybe not right now.

DR. STRAND: Okay.

CHAIRPERSON ABRAMSON: Yes?

DR. SIMON: Dr. Woodcock has something that she might want to add.

DR. WOODCOCK: Well, I don't want to

interrupt the flow. So go ahead. You know, I want to talk about the claim you're talking about, you know, at some point.

Go ahead.

CHAIRPERSON ABRAMSON: Well, I think maybe we can do that. I just wanted to close out Question No. 2, and then we could go to Question 3, which I think begins to address that.

If that's all right, we'll have Dr. Siegel make his presentation as well and then get into that issue.

So with regard to Question No. 2, it sounds like the consensus of the committee is somewhere between six and 12 months is a reasonable duration of a randomized trial from which you ought to be able to see meaningful and sustained responses in the HAQ disability index.

If that states the committee's -- so I guess for Question No. 2.

All right. Now, what type of data are needed to assess durability of effect beyond an initial superiority study period?

Perhaps, Dr. Woodcock, perhaps you can make your comment here, and then we can get into this discussion if that's all right.

DR. WOODCOCK: Certainly. As I said, I don't want to interrupt the flow, but I think when we wrote the initial guidance and had the discussion of disability, we were talking about something different than what you're talking about here today.

In here you're talking about a measure that's fairly responsive, as we found out, as Jeff was talking about earlier, to these newer therapies in a fairly short amount of time.

And so the claim, if you write a claim that is just improvement in physical function, that is a symptomatic claim basically, right? And you know, so the amount of time to demonstrate that claim really relates to number one: how fast does the agent work, which was already raised, okay, and how long do you need to observe to see that, combined with what is the clinically meaningful duration of improvement in that symptom of diminished physical function?

And I think that's quite different than the notion of progression of disease over time, which is something that was really wrapped into that guidance originally. I would just like people to keep that in mind.

CHAIRPERSON ABRAMSON: Yeah, it is an important discussion. I think Dr. Wolfe even began to address that, too, about what it is that we're talking about that is function that isn't picked up in some of these pain domains.

I don't know. Dr. Fries, do you want to comment on that?

DR. FRIES: I don't have too much to add, but it's obvious that when you take a bunch of different things that are supposed to measure either process or outcome, number of tender joints, number of swollen joints, physician global, patient global, HAQ disability, and so forth, that you see in almost all of the results that they move in parallel. Some are more sensitive than others, and some are conceptually superior to others in terms of saying what it really is that we want to say.

But it shouldn't be surprising that they are imbedded in each other, and it would be surprising if the number of tender joints weren't associated with pain and the pain weren't associated with disability, and the dissection of how much disability is caused because you are not able to do it because it hurts too much versus you're unable to do it because your joints are too stiff or some other kind of reason.

To me we're after the greatest sensitivity and the greatest kind of clinical and human relevance that we have, and it's in that area that I seriously want us to move toward looking at disability or improvement in physical function because it's more than a symptom.

It's sort of a symptom, Janet, you know. I mean, that was sort of what I was trying to indicate before. Pain I'm pretty sure is a symptom, and so it's a complex measure which reflects a good hunk of what the patient wants, and as such, I find it justified.

DR. WOODCOCK: Could I response?

CHAIRPERSON ABRAMSON: Yes, please.

DR. WOODCOCK: You know, I'm agreeing with you. I'm simply saying as far as the duration that you need to observe improvement in that particular measure, all right, is it's more like symptoms than it would be long term functional debility or whatever you want to call it because it's very responsive.

And so the question really is, and, Lee, you can correct me, but when you construct a claim about that, how long so you need to observe improvement in that measure before you're convinced that the patient has improved in those measures, which we all, it sounds, agree are more globally meaningful than simply measuring the joint counts or whatever.

That's all I'm saying, and I think that's really the task if you're talking about revising the guidance, is simply saying how long do you need to observe improvement in that measure or whatever, change over placebo or active, until you're convinced that there has been an improvement

in whatever is measured by that measure.

CHAIRPERSON ABRAMSON: And there the question really is how long can you sustain a placebo controlled trial versus how long you need to be sure the effect is maintained over time after the ending of a randomized trial.

DR. WOODCOCK: Well, how long -- I would leave aside the placebo controlled trial first because that's a problem. How long you as rheumatologists would want to observe your patient to be assured, using the HAQ, that they'd had a clinically meaningful improvement on the HAQ, right?

Yeah.

CHAIRPERSON ABRAMSON: Jim, Dr. Gibofsky.

DR. GIBOFSKY: But Dr. Woodcock's comment raises another interesting dilemma, and that is the difficulty of extrapolating clinical trial data to clinical practice and the observational methodology that we use at the conclusion of a clinical trial with its inclusion and exclusion criteria and the metrics that we use to follow up

patients thereafter, I suspect you would find that they were not as precisely calculated, but go more either with a sub-analyses, perhaps a physician's global assessment, rather than the precise things and multiple subcomponents for use in a clinical trial.

So I think somehow we have to get at the dichotomy when we extend beyond the clinical trial period for continued maintenance of what instruments are being used in clinical practice.

CHAIRPERSON ABRAMSON: Dr. Goldkind.

DR. GOLDKIND: Yeah. Getting back to the databases that were presented that deal with this issue, it appeared that there was separation from placebo early on, which at least answers for this product that it's a fairly early phenomena that there would be benefit in as picked up by the HAQ instrument.

And then the issue of durability. Do you believe that it's a sustained benefit? Number one, you want to be sure that you're not missing simply a lag in the placebo group. Maybe they would

have improved at month two and you've defined month one as the endpoint of observation, but it did appear that whatever effect placebo had, whether you looked at it, I believe, the LOCF or the completer analyses, you got to a level of stability quite early after at least the three month time point.

Now, whether we looked at the monthly HAQ, you know, there may be a little bit of noise in there. I don't know whether it's three months or four months, but once you did establish what the placebo response was and what the drug response was, it appears that that was stable over time regardless of the analysis.

CHAIRPERSON ABRAMSON: We should move on to Question No. 3. What type of data re needed to assess durability of effect beyond an initial superiority study period?

And, Dr. Siegel, is this the appropriate time for your presentation?

DR. SIEGEL: I just wanted to say a little bit from the analysis of the Remicade data on HAQ for two years, some comments about these

different analyses. There is obviously a tension between trying to get complete ascertainment at the two year time point and the problem that patients who failed to have an adequate response tend to drop out, particularly as the patients in the placebo arm.

In that regard, in the Remicade database we had 70 percent HAQ measurements at two years, which made it very helpful for feeling that there was a fairly complete analysis of the data and slightly higher percent of the Remicade treated patients with HAQ measurements at two years.

For some of the reasons that have been discussed, we were uncomfortable with relying too much on the last observation carried forward. For one, it's content to over estimate the treatment effect because patients who drop out early who would have deteriorated over the two years might be counted as having a good response whereas they might not have had they stayed in.

So what we have used instead in many of these studies is a non-responder imputation. This

allows you to maintain the intention to treat analysis, but you look at the analysis a little bit differently. You look at it more as success or failure of therapy with respect to the endpoint that's being looked at.

So with respect to the HAQ, you would consider anyone who dropped out before a certain time point as failure of therapy, but anyone who had an improvement of a certain level or great and that was maintained would be considered a responder.

So the specific analysis that we as sensitivity analysis for the Remicade study was to look at the minimal clinically important differences determined by studies by George Wells and others of .22 units of improvement. We chose an amount slightly higher than that of .3 and considered patients who had an improvement of .3 or greater at six months and 12 months to be responders for the one year end point and for the 24 month endpoint we considered someone a responder if they had an improvement of .3 or greater at six months and 12 months and 18 months and 24 months. Anyone who did

not meet that level of improvement or dropped out was considered a non-responder.

And we saw significant improvement with these non-responder imputations with these responder analyses. It gave us some comfort level that the improvement was real.

And I just want to mention that all of these analyses were included in our briefing document of yesterday covering the safety and efficacy of the TNF blocking agents

CHAIRPERSON ABRAMSON: Thank you.

Any questions for Dr. Siegel?

(No response.)

CHAIRPERSON ABRAMSON: All right. If we go to Question No. 3 what I'd like to do is ask Dr. Elashoff and Dr. Anderson to first respond to this question and Dr. Makuch, if they wouldn't mind.

What type of data are needed to assess durability in terms of maintenance of effect size seen during initial superiority study in ITT?

If you could look at this question, and from a biostatistical perspective give us your best

insights.

DR. ELASHOFF: Well, Dr. Siegel basically just said that in the case of a previous approval they did not, in fact, analyze it in a way at all similar to what's been analyzed here today, but made certain definitions of what's a responder and what's a non-responder that people ended up feeling comfortable with.

The whole issue of maintenance of effect size basically requires you to continue to have two groups to compare and then some comfort that the size of effect that you're measuring has not been influenced too much by missing data issues and so forth.

I would like to support the idea of alternative approaches to the analysis like the one that Dr. Siegel talked about or like the one that Jennifer Anderson was talking about where you actually, if possible, actually got measurements at the end of, say, two years for everybody no matter where they had gone in the meantime and talk about whether the ones who had started on your drug were

better off at the end of that two years than the ones who had started on something else.

But any kind of attempt to sort of keep measuring a difference as you go along with lots of people dropping out is problematic on the face of it.

DR. ANDERSON: Well, actually if you're really just asking about durability of effect and you found an effect in the randomized trial, say, in six months, it would seem to me that you can do an analysis of the stability of the effect in just -- you know, even if you lose your placebo group at that point, you can still continue with the patients with the active drug and look at how stable that is.

Of course, you know, probably people think of reasons that that's not adequate, but on the face of it it seems to me it might be as long as you really had a good placebo controlled, you know, or comparator controlled six months randomized part of the trial.

DR. WOODCOCK: I wonder if it wouldn't be possible. Obviously these kind of analyses that

have been presented today were specified by the FDA, and that's the way they've asked the data to be analyzed previously, but wouldn't it be possible to construct a separate endpoint after termination of the first part of the trial, which would be a sort of kind of survival analysis where you define failure and the survival analysis would be ability to maintain a certain level of whatever function?

And then you would look at whether they dropped out because of side effects or loss of efficacy. You would just look at the survival analysis subsequently.

DR. MAKUCH: I guess I'll respond to that. I think it's a good idea because, again, I see different issues here. I'm being very literal when I look at the what I think of as being durability of effect, and so I think then to me it opens up potentially different endpoints to be considered, and I think the endpoint that you mentioned would be at least one to really look at.

Durability and then trying to pick A, B or C here from Question No. 3, actually I guess I

would pick none of them. The reason is because effect size to me means the difference between an active drug and some other drug or placebo, and to me durability effect, I think as Dr. Anderson was saying, is really just if an effect has been established at some period of time conditionally on that group, then is it maintained; is it durable?

And to me then it does just get at is there stability. One can then even with missing data look at the trajectory of each subject over time. So if they don't go out to the entire two year period, let's say, from six months or one year when the effect has been established, then from that point forward you can measure either with the scope or some other kind of situation for each subject individually so they don't have to get out to two years some trajectory and indication of stability.

So that to me is what the durability means. The effect size, which to me means a between group comparison, does not really enter into that equation, but then it gets back to the other issue of what is the hypothesis. Is it durability of

effect, which to me means conditionally that you did have an effect; what's the trajectory for versus the other hypothesis, again, which is sort of being floated around, but I'll try to be more focused and say that the other one is using an ITT population, and then going from zero, let's say, out to two years. You could still use all of these other analyses, but that to me is a very different hypothesis than the durability of effect.

So I think, number one, you have to decide what are you really -- which hypothesis are you really interested in? I think then it would drive what group of people you look at, what the methods of analysis would be, and perhaps what alternative endpoints would be considered.

CHAIRPERSON ABRAMSON: Dr. Gibofsky.

DR. GIBOFSKY: I agree with that. I think the other problem that you raised before is this issue of what are we looking at. Are we looking at a difference between zero and two years or a difference between one and two years? What is the trajectory?

And I'm struck there because, as Dr. Choi told us, that where one has missing data early on or at a certain point in time such that you're imputing the next point, then you're basically bootstrapping to go forward on imputation of data that was missing to begin with.

I wonder then to what extent we should be asking not just what type of data are necessary to assess durability, but what kind of methodology should be applied to that data, as Dr. Elashoff has suggested, in order to be convinced that what we measure is, in fact, reliable.

CHAIRPERSON ABRAMSON: Dr. Simon.

DR. SIMON: Just for clarity, since I am not a biostatistician even in my worst dreams or nightmares, it seemed to me, Bob, that the presentation that the sponsor gave kind of gave the kind of presentation that you were suggesting about durability response in that they measured a response at some point in the first year. There was some issues about LOCF in the first year, but in the second year by taking a year two cohort, which was

only those patients then in that second year, they demonstrated a manifestation which showed that the HAQ continued to respond. I can't remember the percentage, but it was in a high percentage of patients.

Would that be the trial design that you're thinking about in the context of maintenance of response?

DR. MAKUCH: In general, yes. I thought that -- I haven't complimented on the clarity of their presentation this morning, but I guess I will do so now, but, yeah, for the durability of effect, that is to me what durability means. I mean, we can discuss later on some specifics of what they did, but in general, it is conditional that you do have an effect at one year and then how you proceed forward and what happens in that subsequent period of follow-up.

CHAIRPERSON ABRAMSON: Dr. Strand, do you want to comment on the ITT and the durability of effect?

DR. STRAND: I would like to do that,

yes, as we can also show you the analyses around the percent of patients who achieved MCID, which is essentially what I think Dr. Siegel was pointing out.

And, again, I'm reminding you that we were looking for durability of effect because we're talking about studies which maintained their blinds for a two year time frame and then had continued extensions which were also blinded.

If I could have Slide 186, please.

We understand we're comparing two different studies here, but you're seeing on the left the ITT population at 12 months, and you've seen those numbers before, but you also see the percentage of patients achieving MCID, and you're seeing on the right the year two cohort for US301 and 85 percent of those patients completed a full 24 months, and you're seeing that the same percentage of patients had achieved MCID in both of the active treatment groups.

If we go to the next slide, you see a similar type of analysis for the six month that was

carried to 12 months. It says 12 months, but it's six months for the MN301 study, and then the year two cohort, and again, we're talking about those patients who entered the year two cohort, obviously a small number of patients, but it's a maintenance of effect, and the percentage of patients who achieve MCID is either increased or the same.

And if we go to the third one, again, I'm showing the similar type of data.

Now, if I recall, the ATTRACT trial was actually unblinded because an IRB stated it was no longer ethical to keep patience on placebo some time around 12 months. So that I can only understand the 102 week data in the context of that, and as I'm saying here, yes, the placebos have all been exited, more or less all been exited from most of these studies, but these patients continued to be blinded as to treatment.

Final slide.

So this is just another analysis to try and look at what we call a response and clinically meaningful improvement in HAQ disability index to

point out that the patients who achieve MCID in the first year are usually the ones who continue to have that response in the second year, suggesting that the people who go from yet to no are only nine and five percent in the two active treatment groups.

And finally, I know that Dr. Cook has had a lot of thought about LOCF analyses and a lot of discussions with us about durability of effect in these studies, and I wondered if you'd let him just speak briefly.

DR. COOK: Gary Cook, Biostatistics Department, University of North Carolina.

I think one consideration that you should take into account in these discussions is that when patients drop out, you sometimes have different types of information on them. If a patient drops out for lack of efficacy, it may be more reasonable to do carried forward because had they continued on the treatment that had given them lack of efficacy, they may well have continued to get worse.

The patients that are more tricky to

judge are those who discontinue for other reasons, like adverse events or just simply it was not convenient for them to stay in the study.

But in these cases, the vast majority of patients, particularly in the placebo group, did drop out for lack of efficacy, and I think that kind of information can be fairly helpful.

With respect to the question of durability, I agree with some of the points that others have made, that if you establish by intent to treat type analyses statistically significant differences at an early time point, like four months or six months or possibly one year, that addresses the efficacy question.

For durability, in my interpretation, there's sort of two components that are important. One is that a substantial fraction of the patients who completed one month -- I'm sorry -- 12 months are still there at 24 months. So usually you would want to say that at least 80 to 90 percent, maybe more than that, of the patients who completed a 12 month visit are still there at 24 months because if

you had large numbers of people dropping out between 12 months and 24 months, then whatever you saw at 12 months might not any longer be durable.

And then to the extent that you have data at 12 months and 24 months within a particular group you'd like to see relatively small change between 12 and 24 months. There are some ways of trying to statistically quantify both of those. We've been in this discussion more or less just talking about principles for them, but I think durability does have both of those components, that between 12 and 24 months there's relatively few dropouts to support durability, and also for those patients that have real data at both 12 and 24 there's little change.

CHAIRPERSON ABRAMSON: May I ask you, Dr. Cook?

DR. COOK: Oh, sure.

CHAIRPERSON ABRAMSON: I just had a question. I guess from Dr. Choi's analysis the concerning point to all of us perhaps is that only 28 to 30 percent of the people who were sustained

and followed at the two year time point, and that a lot of the statistical difference between the leflunomide and the sustainability was due to the patients who were the last observation carried forward, which represented about 70 percent.

So how do we think about that, that the statistical significance may have been done as a result of the imputed values of people who are no longer in the protocol?

DR. COOK: Well, the first thing you have to recognize is that the observed case analysis that he displayed has to be looked at very cautiously, particularly for the placebo group because the placebo people who continue beyond 12 months through 24 months are all patients who are doing very, very well, on placebo and is a relatively small fraction of the group originally randomized to placebo.

Secondly, as I said, there are two types of missing data. There are people who discontinue for reasons of lack of efficacy, and for them last observation carried forward may well be optimistic

because those are the patients who you could argue you should carry forward the worst possible value.

And then there are other people who discontinued for unknown reasons or reasons unrelated to efficacy. Those are the ones for whom the results from last observation carried forward might need support from a variety of sensitivity analyses.

Some analyses would say suppose that they would have responses in the future like placebo patients. Others might basically say that you would give all of them the worst possible value.

But I think you need to recognize that in the placebo group the individuals who discontinued placebo for lack of efficacy, and this would similarly apply to the other groups as well, any patient who discontinued for lack of efficacy really should be either given the last observation carried forward or the worst possible value.

And if you were to do analyses looking at the data that way, you probably would see a picture not all that different than what the

original LOCF analyses did.

The people who drop out for lack of efficacy are called informative dropouts. They drop out in a manner in which you sort of know what their status was at the time of dropout, and for them it is reasonable in many cases to say the carried forward value is a fair value to use for them.

It's the people who drop out for other reasons that have all sorts of uncertainty.

DR. STRAND: Not wanting to be difficult, but I can actually show you the slides of the dropouts over 24 months and the two active treatment groups so that you can see what's happened to the HAQ.

CHAIRPERSON ABRAMSON: But not just yet. I'd like to hear more from the committee.

Dr. Elashoff.

DR. ELASHOFF: With respect to the issue of durability as we were talking about it, which has to do with change between 12 and 24 months, I don't think we have actually seen that data because I think everything we've been shown goes back to

baseline again.

So aside from the slide that had the yes/yes and the no/yes, and that came by pretty fast and I didn't know how dropouts were handled in that respect, I don't think we have actually seen today the direct analysis of change from 12 to 24 months, and certainly even interpreting that we would need to know what's been done about the dropouts and how worried we are about how many there were.

DR. MAKUCH: Two remarks. First, I guess, responding to Gary Cook, I think again it goes back to the question if you're looking at the conditional at 12 months, I like to design away my problems as much as I can and so therefore if you look at the conditional 12 months, then maybe that's one way to get rid of everything that happens in the first year.

And secondly, I think Dr. Choi did it from the start going out through two years, and I think the problem has become more magnified as you go further out.

I guess responding to Dr. Elashoff and

getting at the data, I actually do believe that the two year data conditional at one year have been presented. For example, at Slide 60 the HAQ then is presented where it is, the two year cohort at 24 months. I believe that that is based on the information at the end of, let's say, year one and then conditional at year one going out to year two.

DR. STRAND: That is correct. Every year two cohort is defined as patients who enter year two, have a visit after month 12, on or after month 12, and it's ITT from month 12 to 24, and again in all of these treatment groups, the dropout rates are on the order of ten to 15 percent.

DR. ELASHOFF: So the baseline here is the one year baseline?

DR. STRAND: No, it's the two year baseline.

DR. MAKUCH: Well, it's the start at year two and then the end at year two. But my question about these slides are, in fact, if you leave that one out, for US301 you start out with 97 people and 101 people in the two treatment arms

respectively, and if you then go back to Slide 41 where you do have what you call your Y2C, your year two cohort, you do, in fact, have 98 and 101. So, therefore, that's the start, 98 and 101.

And then the slide that preceded this one that you just showed going to 97 and 101, that follows closely. My problem actually is so to me it is conditional at year one. Then what's happening in the year two period.

But my problem is with the subsequent two studies, unlike US301 where you did have that kind of comparability between the baseline or the numbers in, let's say, Slide 41 or Slide 44 or 45 for the 301 or 302 studies, it does not then carry over the number at risk at the start of year two, does not carry over to these numbers that you see here, unlike the very nice correspondence that you do see with US301.

So my remark is there were fewer number in MN301 and MN302 than there should have been. The number of numbers that you have in US301 are appropriate based on Slides No. 41, 44, and 45.

So I guess I need clarification because I agree with your conditional results for US301. It is a subgroup that you're using for the other two studies.

CHAIRPERSON ABRAMSON: Do you want to respond to that, please?

DR. STRAND: Yes. The clarification is that the other two studies were extension studies, and we have the reasons that patients chose not to enter those extension studies, and that's why they were lost.

And we have actually more detailed analysis than this, but I'll show you this one slide and that should make some of the point, and that is you see the patients who choose not to go into extension MN303. Of the 16, ten and seven, they are divided between those who are actually responding at that six month end point and those who are not, and the same analysis goes forward to the MN305.

DR. MAKUCH: But let me ask you a question because I actually will respectfully disagree.

DR. STRAND: Okay.

DR. MAKUCH: I'll look at your Slide 44.

DR. STRAND: Okay.

DR. MAKUCH: And so when I look at your Y2C, which is the number at risk starting at year two, you have 60 in each of the two arms, and that's what I thought would have then be carried through in the previous slide that you showed for the MN301 data.

Because if you look at your Slide 41 --

DR. STRAND: Yes. We have a smaller number. You have a good point.

DR. MAKUCH: And go to Y2C in Slide 41. You see 98 and 101, and then as you go down to your results, you had 97 and 101 for your conditional year one to year two results. It, therefore, corresponds nicely to Y2C.

The Y2C though does not match --

DR. STRAND: Correspond.

DR. MAKUCH: -- between Slides, I guess, 44 and Slide 60.

DR. STRAND: And the reason there is

that there were a certain number of patients in MN301, 303, 305 who did not have HAQ disability index because there was no adequate translation into their language.

DR. MAKUCH: Fine. So I then want to point out that then for MN301 and MN302 and the subsequent follow-up studies that were conducted, that the numbers that were presented for those analyses do not correspond to the Y2C because of the missing data for HAQ unlike 301, in which the number at risk for that conditional analysis for US301, in fact, I guess, must have had except for one patient all of the HAQs, and therefore, it's a more complete analysis based on the number at risk at the start of year two.

DR. STRAND: You are correct.

DR. MAKUCH: Okay.

DR. STRAND: And we do have an unfortunate problem about the HAQ and MN301, but in MN302-304, that is simply what we have.

DR. SIMON: This has been a wonderful discussion for us. We've heard all of the comments

about the issues associated with durability of response, but I just want to be clear that we're not looking for an indication of durability of response. We are just looking for advice on how one would reconstruct this particular indication within the guidance document to insure that we're conveying the most useful information for clinicians and patients to understand after we decide on whatever the primary endpoint is going to be what subsequently happens in those patients.

And I think that Dr. Makuch's clear observation of a response period and a second maintenance period, and then using perhaps this example that we've just seen today as an example of how one might go about that is adequate for us to be able to move on.

CHAIRPERSON ABRAMSON: Shall we move to the fourth question then?

Are the data on leflunomide presented by the sponsor adequately robust, effect size and robustness of database, to support labeling for improvement in physical function?

Who would like to?

DR. MAKUCH: I'll make one very brief comment. I actually do like the conditional analysis that the company did. I thought it was clear, and except for some of the missing data pointed out, I really think it was a very nice way to go.

CHAIRPERSON ABRAMSON: Dr. Elashoff.

DR. ELASHOFF: I still have a question about that because Slides 59 and 60 show changes which -- and it says baseline. It doesn't say from 12 months, and if you look at the Slides 57 and 58, they don't show any change from 12 to 24 months. So those differences should be about zero with some standard deviation.

So either this baseline on Slides 59 and 60 really is baseline and not the 12 month starting point, in which case they don't have the analysis you were talking about, or I'm really confused somewhere.

DR. STRAND: We did two different analyses. This analysis is mean change from

baseline -- next slide -- and it's showing it's the year two cohort at 24 months, and those baselines are 12 month baselines.

We then showed, although we were not comfortable as saying that that was the primary analysis, what happened in the year two cohort since they were in for zero to 12 through 24 what their changes over time were, and that's why you will see there they're going back to the original baseline.

But I'll let Dr. Hurley explain it sine he's been the statistician on this project.

DR. HURLEY: To be clear, this slide shows the change from the original baseline in the year two cohort at 24 months. We also showed the data for the same year two cohort at 12 months and 24 months and showed that those were the same.

So that there, indeed, was no change from 12 to 24 months in the change from the original baseline

DR. FRIES: Just to indicate that I've had a little worry through the morning about the tyranny of the MCID. We could actually leave that

up there because this is itself a subject for a full day, but I just wanted to give a couple of comments because that 0.22 is a line not drawn by patients. It's drawn by health care researchers as being a minimum clinically important difference, and if you actually ask patients, all other things being equal, will you accept a very small improvement, they'll say yes. So that itself it's a little bit of a funny construct.

Secondly, as we're moving from an era in which the average RA patient has a 1.2 HAQ DI to one in which they have a 0.8 DI, the percentage required by the MCID as an absolute value in an area where proportionality may be more important than absolute changes to get around some of these things is going to get us in trouble with the next generation of drugs.

I don't think that it's terribly relevant to this right now, but sooner or later we're going to want to accept drugs that have a marginal benefit of less than 0.22 as being clinically important additions to our armamentarium.

CHAIRPERSON ABRAMSON: Okay. Thank you.

Other comments from the committee members?

(No response.)

CHAIRPERSON ABRAMSON: Why don't we perhaps go around the table and address Question No. 4? Are the data robust enough to support labeling for improvement in function?

Shall we start at the end of the table there? No?

DR. SEEFF: I don't think I should. I'm not a rheumatologist.

CHAIRPERSON ABRAMSON: All right. Abstain.

DR. LEWIS: The only question I would ask is with the infliximab data, I wasn't here yesterday to hear it. How many dropouts were there in that study? Is it comparable to rheumatoid arthritis with this drug?

CHAIRPERSON ABRAMSON: Those issues weren't really addressed yesterday.

DR. LEWIS: Do we know an answer? Were

you left with 25 or 30 percent of the patients in the trials?

DR. SIMON: It's not -- we have the answer, but the answer is not applicable to this particular trial because they're entirely different designs, and because of the issue of the short term placebo exposure, the fact that it was blinded over two years and not the same as the ATTRACT trial, it doesn't even help us even understand that. That's the problem

CHAIRPERSON ABRAMSON: Dr. Day.

DR. DAY: The data presented this morning seem to support number four. However, whatever we decide or the agency decides about number three, our views may change or be modified somewhat.

(Laughter.)

DR. FRIES: I had already said yes on the slide, and I'd give it a higher level of confidence because of all of the studies that have been done with the HAQ over time which show that the best predictor of future HAQs are present HAQs, and

that, in fact, about 70 percent of the variance is explained by the prior HAQ levels.

So this suggests to me that it's very, very likely, and I showed the other slide to indicate the same thing, that there will be durability if you can document the initial response as substantial.

DR. BRANDT: Also, yes, I was initially very concerned about the missing data. The discussion has helped clarify that, and I think that the improvement is real in the initial period and sustained in the 12 to 24 month period.

DR. ELASHOFF: Okay. I'm going to distinguish between the possibilities for what the data might or might not show and the analyses that we actually have in front of us today, and it depends on exactly what time point you choose to say whether you've seen some superiority or not.

Probably if you picked the six month period and really looked into the missing data appropriately and assured us that there wasn't too much last observation carried forward for that data,

that might well be robust enough here.

I have still not been convinced that I have seen what I would need to see for the duration question if we were going to talk about what's happened between 12 and 24 months. It seems fairly stable, but I would want to look personally at the data and at a different way than it was looked at here.

So the data themselves might be good enough if I could see the analyses that I needed to see, which I haven't seen in enough detail today to feel comfortable about.

CHAIRPERSON ABRAMSON: Dr. Makuch.

DR. MAKUCH: I do find that the data are consistent with a claim for improvement in physical function. I do share the concerns though of Dr. Elashoff, and I think as you move forward and take into account previous remarks depending on precisely the time point that you're looking at, depending on the precise nature of the dropouts, I think you heard very excellent remarks from Dr. Cook as well that, you know, more work is needed.

But I certainly see that it's going in the right direction, and they certainly are consistent with this claim.

CHAIRPERSON ABRAMSON: Dr. Anderson.

DR. ANDERSON: Well, I would answer yes. Although the words in parentheses defining robust, "affect size and robustness of database," I don't think apply because what I'm answering yes to is the durability of effect rather than all of these things about effect size, which I don't think can be really adequately answered given all of the dropouts in the latter part of the trial.

CHAIRPERSON ABRAMSON: Ms. McBrair.

MS. McBRAIR: Based on Dr. Fries' comments, my answer is yes.

DR. WILLIAMS: Yes.

CHAIRPERSON ABRAMSON: Well, I certainly accept the effect on the HAQ disability index at the shorter time points. I'm a little concerned about Dr. Choi's analysis, although I'm not sure that I have enough data to talk about a two year endpoint to be absolutely comfortable with that.

So I have some ambivalence about whether more data would be necessary at the extension period

DR. MANZI: First of all, I certainly have gained a lot of insight in how many different ways you can look at data.

(Laughter.)

DR. MANZI: But let me just say that I think it was an incredibly good discussion, very fair discussion, and a term I'll use from one of my colleagues here is this idea of conditional analysis where you take people who have clearly made some predefined effect size difference, and then is there durability beyond that point I think is a fair way of looking at it.

My only question that I don't think we've addressed is how many people or what percentage of the original cohort would you accept as being a clear representation of durability, and maybe this isn't fair, but if you start with 500 people and you get a response at 12 months, and then you want to look at durability, if three of those people remain in the study and their response is

sustained, is that a legitimate -- is that durable? Yes, that's durable, but does that represent that this drug has durability for the majority of people that you use it on?

And I think that's the question that we're grappling with at least in my mind.

Anyway, I also like the idea of perhaps deciphering a little better the imputed cases because there's different reasons as was pointed out for withdrawal in the placebo group, some where you feel more comfortable potentially carrying forward and others not, and maybe some additional looks at those imputed cases on that stratification may help.

You're going to force me into a yes or no. I'll say yes.

DR. GIBOFSKY: I agree with Dr. Abramson. It was a concern for me, and I'm weighing the notion of the high rate of missing data and the validity of the two year analysis with imputation of the year one data. That creates one problem, but on balance I accept Jim Fries' notion about the tyranny of the MCID.

So overall I would say the answer is yes, but I would retain the right to change that, as Dr. Day pointed out, if the definitions in Question 3 which are changed.

CHAIRPERSON ABRAMSON: Okay. Thank you very much.

We unfortunately are running late. We'll break for lunch, but I'd like to ask people to be back by ten after one so we can get the afternoon session started.

Thank you.

(Whereupon, at 12:35 p.m., the meeting was recessed for lunch, to reconvene at 1:17 p.m., the same day.)

AFTERNOON SESSION

(1:17 p.m.)

CHAIRPERSON ABRAMSON: We're going to begin this afternoon's session with another open public hearing, and the first speaker this afternoon will again be Dr. Sidney Wolfe.

Dr. Wolfe.

DR. SIDNEY WOLFE: The first two minutes of what I have to say may not immediately be apparently connected with this topic, but it is.

Five years ago we did a survey of medical officers in CDER and found that a number of them felt that their views were being suppressed in terms of participating in FDA Advisory Committees. As I remember there were 14 instances they cited where they were told not to present information at FDA Advisory Committees that was unfavorable to the possible approval of a drug.

CDER itself did a study two years ago because they were concerned about the tremendous turnover of highly trained personnel, physicians and others in CDER, and they found about a third of the

respondents didn't feel comfortable expressing their differing scientific opinions. Over one third felt that their work had more impact on the product's labeling and marketability than on public health.

And the recommendation, and this is quite relevant to what has happened at this meeting, the recommendation from the FDA was to, quote, encourage freedom of expression of scientific opinion.

Dr. Woodcock, I think, very correctly stated that there was a sweatshop environment, end quote, that had come upon CDER since the Prescription Drug User Fee Act of 1992. I think that is absolutely correct.

Unless this openness occurs, and today, as I will mention, is an example of where it wasn't, the best people are going to leave the FDA. We have three former CDER employees on our staff half time now, and it is in no small measure due to this kinds of problems.

The concept of generating a signal from adverse drug reactions is a very important one.

It's why all of the energy is spent collecting the information, processing it, and many people in the FDA first and foremost looking at it.

But it's not going to make a difference if the signal isn't taken seriously and the action based on the signal isn't prompt and appropriate to the strength of the signal, especially when the signal, and it has happened too many times, confirms a signal that was already there from randomized controlled trials on the same drug. Troglitazone is an example of that. Rapacuronium is an example of that. There are a number of examples, and I think this is another example.

In too many instances serious post marketing safety problems identified by the Office of Drug Safety have not been acted upon because of resistance of FDA management and from the Review Division that originally approved the drug.

An extremely thorough review of the hepatotoxicity and other problems, including the discussion, a very good discussion, of possible risk management strategies was done upon request by Drs.

Banelle and Graham in the Office of Drug Safety and signed off on by Dr. Beitz, the Director of the Division of Drug Risk evaluation, despite this 37 page evaluation which concluded the drug should be withdrawn from the market.

None of the authors of this review were allowed to present their work to the Advisory Committee and to be questioned by you in terms of what you agree with, what you disagree with, and instead, a much in my view less thorough review by someone in the Drug Review Division, Dr. Goldkind, who is in the Drug Review Division. He's not in the Post Market Surveillance Division -- will be presented that in my view attempts to whitewash the findings of the Banelle-Graham review, another blow to scientific morale at the FDA and another example of the Review Division sort of riding over, in a sense, the post marketing surveillance people.

I'm going to mention a few things from the reviews by Drs. Banelle and Graham and then just weave in a couple of things that you may not have noticed that were in our petition that we filed a

year and a half ago to take this drug off the market because of its hepatotoxicity.

The Banelle-Graham review identified 16 cases of leflunomide related acute liver failure, 12 probable, four possible, and 38 cases of leflunomide related other severe/acute liver injury.

The monthly reported hazard rate for acute liver failure and for other severe liver injury appears to remain relatively constant with continued use of the drug, and a term which others have used, which is the number needed to harm, as in the number of people needed to cause harm, range from 107 to 188, a mean of 150, at 23 months of continuous leflunomide use. And that's harm as in acute liver failure adjusted for under reporting. These risks are extremely high.

One of the things which I had not seen until yesterday when these data were at least put up on the Internet, even though they're not being presented, was the extraordinary fall-off in people using leflunomide. A database made up of three and a half thousand patients from Tennessee Medicare or

Medicaid, rather, and TennCare and the United Health Group -- these are organizations that the FDA routinely contracts with to look at patterns of use and possible patterns of injury of drugs -- these data showed that by four months half the people that started on this drug were no longer using it.

The median duration of leflunomide use was four to five months with 19 percent only continuing for greater than a year and only six percent of those starting to use it continuing for greater than two years, less than one percent for greater than three years.

These are data from 1998 through 2002.

In contrast, certainly we know that there is some kind of hepatotoxicity with methotrexate as well, but the methotrexate as used in rheumatoid arthritis is not associated with severe/acute liver injury, which is what we're talking about here, or failure. The main hepatotoxic risk -- and, again, these are taken from the review by Drs. Banelle and Graham -- the main hepatotoxic risk is liver fibrosis. The literature

suggests that the level of fibrosis is usually mild, occurring after many years of treatment and rarely progresses to cirrhosis even after six years of use or longer.

A comprehensive review of the literature on this topic covering 625 methotrexate treated patients with liver biopsies found no cases of cirrhosis.

And as mentioned on one of the slides that Dr. Simon showed this morning, it was a relief back in the late '80s, early '90s when instead of rapidly falling off, as many of the people had with the other modifying drugs, methotrexate allowed people to stay on for a much longer period of time, and again, in this review, they point out that usually up to 82 percent at two years, 76 percent at six years. Again, that's an earlier phase, and it's not with the availability of some of these other disease modifying drugs that are available now.

But it's in sharp contrast to the current in the real world now, extraordinary, I was surprised by, fall-off of use of leflunomide in

those two large databases.

These now are just data from our petition, and it's on the issue of is the signal coming in now confirmatory of earlier problem. Again, these are from the randomized trial, the US301 that you saw a lot of data on efficacy this morning, and this is liver function abnormalities for leflunomide versus methotrexate.

For AST, methotrexate was .5 percent of patients. This is the number or percentage above three times the normal upper -- more than three times above the upper limit of normal for this function. So it's 0.5 percent for methotrexate, 2.2 percent for leflunomide. For ALT it was 2.7 percent for methotrexate and 4.4 percent for leflunomide.

In terms of the withdrawal rates, again, this is a randomized controlled trial. For leflunomide liver function abnormalities, it was 7.1 percent withdrawal rate, which is very high for a study unless the drug is hepatotoxic. For methotrexate it was 3.3 percent. Diarrhea, 2.7 percent for leflunomide, zero for methotrexate.

Nausea, 1.6 versus .5 for methotrexate.

Liver toxicity was also increased significantly when leflunomide was added to the drug regimen of patients who were already on methotrexate and who did not have LFT abnormalities. In Study FO1, the only study to examine this question, 30 such patients had leflunomide added for a period of six months. While taking both drugs, 57 percent had LFT elevations, of which 23 percent were between 1.2 times and two times, but 34 percent of these patients, of the total denominator, had liver function elevations of more than two times, half between two and three and half of them, 17 percent over three times.

Now, these are, again, in people who had already been on methotrexate and who had not had liver function abnormalities at that time.

Going on, why is this drug so toxic? One reason is the extraordinarily long half-life, in a population study, 96 days.

On the other hand, the half-life of methotrexate is three to ten hours. So it achieves

steady state between one and two and a half days.

There's also a lack of a proven effective washout procedure. There have been some little studies, one on one patient, one on not many more, and it's not at all clear that using charcoal or other ways of reducing the amount of drug in the body are that effective.

Pregnancy, another serious concern. We all know that methotrexate is a tartogen (phonetic), and is counterindicated strongly in pregnancy.

We looked at the FDA database. There were no cases reported to FDA of complications of maternal exposure between September 30th, '98, an June 30th of '92. Methotrexate, like leflunomide, has a black box warning. However, for leflunomide, between the end of September '98 and through June 2002, looking at all cases where leflunomide was listed as the primary suspect responsible for the observed toxicity.

And, again, the leflunomide labels had this black box warning since the drug was approved. There were 52 reports of adverse reactions relating

to complications of maternal exposure, including 37 women with either spontaneous or induced abortion, implying either that the label is not being read or that the washout is not effective or, more likely, both because a lot of these people probably don't even know that the problem is so serious you should try washout.

But, again, given that it's not that effective, I'm not sure what difference that would have made.

The last part of the discussion, and again, where I thought it was very well done, but not to be presented to you today except in rebuttal by a series of people, was the discussion of risk management, and again, the FDA has had a fair amount of experience over the last five or ten years on the risk management problem.

A drug gets approved, and some problems occur, in some cases known to some extent, but not as much before approval. What do you do about it?

Duract is one example. It was clear before it was approved, an NSAID, that it caused

hepatotoxicity. It was approved, eventually came off the market because the warning labels didn't work.

Troglitazone was approved. There was some strong suggestion, which we actually asked for a criminal prosecution of the company because of it, there were data showing whopping high liver elevations in the controlled trials which weren't adequately focused upon or delineated.

Again, when troglitazone came on the market, there was absolutely no indication on the label that you should do liver function studies. By the time it came off the market, 12 of them in a year, it didn't work again.

So that when we apply this kind of background problem to leflunomide, the question is: if there's going to be some risk management strategy other than taking it off the market, what would it be and what would the odds be that it would work?

And I think that the answer from experience, particularly in the case of liver

toxicity, is that whatever it is is not likely to work very much because it hasn't worked before.

We do not have an example recently of a label change, that kind of restriction on the use that has worked for liver toxicity. So at best you can say that this kind of attempt would be speculative and unproven.

Again, these are comments made in this very good review of the possible risk management strategies.

The remaining, and I'll read in conclusion from what they said, the remaining risk management strategy market withdrawal is effective at protecting patients against drug induced harm. In our view reliance on methods known to be ineffective, that are experimental in nature, now goes to substituting unproven therapy for proven therapy or withholding proven therapy in the setting of serious of life threatening circumstances.

In the remaining two and a half minutes, I'll again mention what I mentioned this morning. It's really because of the death of a patient from

acute hepatic necrosis by your former co-member of this Advisory Committee and former Chairman, Dr. Yocum, that I became involved in it. I thanked him for this, and as I mentioned this morning, he still is not prescribing this for his patients.

It is entirely possible to practice good, effective rheumatology without the use of this drug, and we still hope that the FDA with or without your advice will realize that it needs to be taken off of the market.

Again, it's a matter of no unique benefit. The one large trial which, yes, did not have folic acid and, therefore, it's not valuable as Dr. Simon pointed in terms of looking at liver toxicity, and the looks at liver toxicity that I just mentioned from the controlled trials did not include that one but the later ones.

But in terms of the effectiveness, that large trial showed that methotrexate was actually significantly more effective. But even if they are the same, which is the gist of what the presentation this morning was, that these are mainly the same in

terms of effectiveness, it has unique hepatotoxic danger, and I hope that it is taken off the market before too many more people are injured by it.

Thank you.

CHAIRPERSON ABRAMSON: Thank you, Doctor.

DR. SIDNEY WOLFE: And I yield the remaining one minute and 15 seconds to the next person.

CHAIRPERSON ABRAMSON: Okay. We thank you. These are important issues, and I'm sure that there's going to be a very fair, open, and comprehensive discussion of each of the items that you've raised.

The next speaker is Ms. Amye Leong, who is a spokesperson for the United Nations Endorsed Bone and Joint Decade.

Ms. Leong.

MS. LEONG: Thank you very much, Mr. Chairman.

And good afternoon to you all and thank you for the minute and a half, Sidney.

I am a public citizen. I'm a concerned citizen. I am what you've been talking about all morning. I'm a person with rheumatoid arthritis. I've been taking, in fact, most of, in fact, all of the drugs that we have mentioned so far and then some.

I have rheumatoid arthritis. I have Sjogren's Syndrome. I have osteoporosis. We didn't know it then, but I started the nation's very first support and education and advocacy groups for young people with all kinds of rheumatic diseases.

I've bene a volunteer with the Arthritis Foundation, have been a volunteer leader and spokesperson for the Arthritis Foundation. I'm a former member of the Advisory Council of the National Institute of Arthritis, Musculoskeletal and Skin Diseases. I'm President of Health Motivation, and in fact, started this company in 1999, the year I actually went on leflunomide, not that there's any correlation, but started a company called Healthy Motivation, which is a health education, motivation, and advocacy consulting firm based in California and

based in Europe.

And so I actually was trying to test out Dr. Simon's hypothesis about the early effects of trying to treat arthritis by doing winters in Europe as well as in California.

(Laughter.)

MS. LEONG: And I can tell you it's not enough.

I'm currently spokesperson for the United Nations endorsed Bone and Joint decade. Many of you have heard of this. The year 2000 and the year 2010 has been declared the decade of the bone and joint, in which there was a focused global attention toward diseases and disorders that affect those of us with arthritis, osteoporosis and other musculoskeletal disorders.

We currently are in 55 countries, including the United States. President Bush endorsed this, and we are now coalescing the many health care professional and patient organizations that work in this area.

But I'm standing before you today as a

concerned patient. This is my very first opportunity to participate in an FDA Advisory Committee meeting in the Arthritis Committee. I'm fascinated by it. I think I've become addicted to it for the last two days.

I have seen that there is, indeed, a great deal of objective review by the FDA, and I look forward to what goes on. Let me provide to you my disclaimers. I understand we as speakers must provide our disclaimers.

My travel expenses from Paris to Washington were in part supported by Aventis to come to participate in an Arthritis Foundation advocacy meeting of which I've been participating in for the last several days.

In addition to that, being here has been an important part of my advocacy, and it was my insistence to be here today.

I've served as a consultant to several pharmaceutical companies, many of which were present yesterday, on nonbranded education items. I have consulted with Pharmacea Aventis, Pfizer, Wyeth. I

provide health motivation speeches which have been funded in part by many of the pharmaceutical companies that have products in the arthritis field.

And so I wanted you to know that I'm standing here today because of the transportation assistance of one company, but most particularly because I am a concerned citizen and a person with rheumatoid arthritis.

The paradigm, we talked earlier about this whole paradigm thing, and my particular case with rheumatoid arthritis and particular experience with it is actually an example of that.

When I was diagnosed at age 18, I was given 18 aspirin, and like all of us who are good patients, we don't question it. We just take it.

Through the years, as Jim Fries so eloquently said, that whole paradigm has shifted to the point where we who are patients have become much more eloquent, much more of an advocate in terms of working with our physicians to understand and ask questions about possible adverse effects.

When I was diagnosed at 18, I did not

know that within six years I'd end up in a wheelchair. Obviously aspirin didn't work.

I spent two and a half years in a wheelchair because I could not raise a fork to my face to eat. I could not walk ten feet. My weight dropped down to 79 pounds, probably the size of some of your dogs at home.

I was truly in Stage 3 severe rheumatoid arthritis. I had recalcitrant arthritis. What you see here standing before you today is as a result of 16 joint replacement surgeries. That's a very, very expensive therapeutic regimen, and I'm still paying for those surgeries at a cost of 25 to $35,000 per operation.

But I'm standing here today because that was the only, only option during that shift of that paradigm.

Today I have been taking and have been on methotrexate for the last 16 years. However, I'm currently on leflunomide, and if I were to listen to my previous speaker, I would think that I would be very concerned.

However, I do not have elevated LFTs, and I'm quite functional, and in all the speaking that I do around the country and around the world, part of my effort as an advocate is to conduct focus groups of those of us with different kinds of rheumatic diseases, and we talk about the three Ds. You know, what is the most important, as Dr. Gibofsky had earlier asked? And certainly all of those Ds are very important.

But most important is the function piece and the discomfort piece. But another piece that you do not address here is the cost piece and the dollar piece. And I can tell you that I am a candidate for many of those, all of those biologic drugs that were presented yesterday, but I chose and I choose today not to be on those drugs yet.

What you don't know is that until there's a cure, I am stuck with a very limited matter of choice. I am stuck with trying to figure out with my physician what is the best possible drug with the least possible adverse effects at the best possible price range for me.

And I know that that is not your purview in the course of your discussion, but those of us who live with it 24 hours a day, it is at the top of our mind because it's either drugs or we eat for that particular day, and that's a horrible paradigm to have to take a look at.

And so I choose to start with those drugs in which cost the least and based on the studies. And I have read all of the information on the Web site and with respect to this particular meeting, and I'm very, very certain that I am on the right course.

Now, when I was crippled, I was very much involved with and very concerned about quality of life. As indicated earlier, i could not function independently at all. I was disabled. I was on disability. I carried that blue card that Jim Fries was talking about, and anybody who looked at me said, "You are disabled, you poor thing."

To have that kind of life is not something that any of us who go into a clinical trial, whether we're on a placebo and we don't know

it or not -- and I'm so pleased that Wendy McBrair spoke up with respect to being on a placebo and having a recalcitrant, serious, painful, lingering disease.

Quite frankly, if you had put me on that placebo, I would have been one of those early withdrawals because I would have insisted the quality of my life is more important than the importance of conducting a trial because it's all about me getting out of pain.

So I can actually understand these numbers. I can understand them with my limited biostatistician background. It makes sense to me.

So function is extremely important for me. Maintaining function is extremely important with the least amount of adverse effects.

I have had all kinds of adverse effects. I've had abdominal pain, fluid retention, gastric ulcers, upset stomach, nausea, vomiting, heartburn, indigestion, ringing in the ears, reduction in kidney function, hair loss, increase in liver enzymes, rash, weakness, unusual tiredness,

sleeplessness, sleepiness, upper respiratory infections, infections, hypertension, elevated blood sugars, insomnia, mood changes, restlessness, diarrhea, constipation, mouth sores, fever and chills, loss of appetite, infertility, missed menstrual periods, high blood pressure, kidney problems, increased hair growth, swollen glands, light sensitivity, bruising, unusual bleeding, weight gain, moon face, muscle weakness, thinning of the skin, brittle bones, cataracts, impaired wound health, hyperglycemia, diabetes, of which I've not had but friends have, osteo, immunosuppression, vasculitis, and these are just some of the side effects of all the drugs that the FDA has so far approved.

I have had those side effects. But yet the risk for me is worthwhile. To me the benefit of having improved function is worth every single one of those adverse effects, and I am willing and most willing to try a drug that provides me excessively relief and particular function.

Sine I've been taking leflunomide nd

since becoming spokesperson for the Bone and Joint Decade, I've been traveling internationally. Ten years ago I could tell you that if you said, "Amye, you have to go to Germany to give a speech," I would laugh at you and say, "How in the world am I going to do that?"

I can tell you that last year I logged in over 140,000 miles, not because of leflunomide, but because of my proactive effort as a patient, as an arthritis advocate monitoring my system, working with my doctor, going in for my monitoring systems of blood tests, having conversations, if not telephone conversations, then certainly by E-mail, so that I am an active partner in my care.

Until there is a cure I am stuck with this disease for the rest of my life. So it's very, very important that I titrate out all of the available options to me, and I'm just glad and very, very pleased that we have an option like leflunomide.

And so I encourage the support of the committee and the FDA to support the sponsor's

request.

Thank you.

CHAIRPERSON ABRAMSON: Thank you very much, Amye.

We're now going to move to a presentation by Dr. Lawrence Goldkind to discuss the presentation of the safety data.

Dr. Goldkind.

DR. GOLDKIND: Thank you.

Larry Goldkind. I'm a gastroenterologist, and I'm Deputy Division Director of the Division of Anti-inflammatory Analgesic and Ophthalmic Drug Products.

I apologize for the density of this presentation and its anticipated duration, and I hope that postprandial sedation does not set in.

(Laughter.)

DR. GOLDKIND: Maybe the blue color will keep us all awake.

Leflunomide was improved in 1998, and at that time, the label did note the potential for hepatotoxicity. To briefly go through the sections,

the cautionary sections of the label as it was approved in 1998, under warnings hepatotoxicity in clinical trials, Arava treatment was associated with elevations of liver enzymes, primarily ALT and AST, in a significant number of patients. These effects were generally reversible. Most transaminase elevations were mild, and usually resolved, although marked elevations occurred infrequently.

And to go on, there is a section within that warning section regarding monitoring of liver function tests and some information on guidelines for dose adjustment and discontinuation.

Also, within the warning section under preexisting hepatic disease, a subsection was established that stated that given the possible risk of increased hepatotoxicity and the role of the liver in drug activation, elimination and recycling, the use of Arava is not recommended in patients with significant hepatic impairment or evidence of infection with Hepatitis B or C.

Under the precaution section, again, the issue of monitoring labs is noted, and also under

the precaution section, a subsection entitled "Drug Interactions." There's hepatotoxic drug interaction caution that states that an increased side effects may occur when leflunomide is given concomitantly with hepatotoxic substances. This was also to be considered when leflunomide treatment was followed by such drugs without a drug elimination procedure.

So that was the state of affairs at the time of approval, and post marketing there have been post marketing reports of hepatitis and acute liver failure, and on the slides I'll refer simply to this as ALF.

These have been received through the adverse even reporting system, which is known to most clinicians as the Medwatch system.

There was a review in 2001 of cases at that time that had been referred. There was extensive confounding and when I say "confounding," meaning other likely causes for liver toxicity in the majority of those cases, and the label is reviewed at that time, and it was felt that the data, the information in those reports was

referenced in the current label.

There was a citizens' petition in 2002 for the removal of Arava primarily based on the reports of ALF, although Dr. Wolfe has outlined some other concerns, and that document is in the briefing background as well for reference.

So based on ongoing concern and reports, an exhaustive, and I emphasize "an exhaustive," reassessment of hepatotoxicity has been taking place of many months now, and that has included assessment of the individual case reports, as well as a reassessment of controlled clinical trials that had occurred prior to approval, in addition to looking at studies that have been done since approval.

And also querying basically any other database that may be available either from sponsor or publications or presentations.

And finally, data mining or an attempt to systematically look at the AERS database has also been performed.

Just briefly to go through, in a sense, the potential sources of safety information in the

drug regulation process, obviously controlled clinical trials is where the safety assessment starts for approval, and I'm going to go through the strengths and weaknesses of each of these in the subsequent slides.

Obviously there are cohort studies, and there's the AERS database. We have multiple sources of safety information. No one of these sources is adequate and sufficient, and they complement one another.

In clinical trials, obviously the strength if that there are comparisons to placebo and as often as possible to alternate therapies so that we have some ability to compare what the different therapeutic options are for particular disease so that physician and patient can be aware as best possible.

These are the least biased. Obviously they're randomized, and so imbalances across groups and channeling bias and confounding factors are minimal in this kind of database.

You get the most detailed information

and you've got the best chance for causality assessment if you do have any adverse events.

And, again, the fact that there are denominators allows you to calculate a rate. Of course, the weakness is for rare events you may not be powered to pick these up, and also exclusion criteria limit the applicability across broad populations. So that for a patient who gets this drug who happens to fit the inclusion/exclusion criteria of the trials, you may have a fair assessment, but for somebody out of the age range, taking other medication or other vulnerabilities, they may not be adequately represented or not represented in clinical trials.

Cohort studies are generally much larger so that there is more of a power to detect events. It's a naturalistic setting, meaning all comers. Hopefully such studies would be done in, in fact, the patient populations that are exposed to the drugs in practice. Therefore, it allows you to identify vulnerable groups and drug interactions.

And it can provide rates for events, and

if you do have comparator groups within the cohort studies, allows some comparative data.

The weaknesses are, again, the fact that these are not randomized studies. It means you've got channeling bias, and so you may have sicker patients or patients who have already been shown to be intolerant to one or another therapy, potentially obscuring differences that may be there in reality.

Causality assessment is a little less robust in this kind of a setting where clinical data may not be available.

Now, the AERS system, obviously it canvasses in a sense the universe of drug exposure in this country, and so hopefully it would have the power to pick up rare events.

And when we speak of the term "signal," it really is most applicable to the AERS database because it does allow you to pick up events that are extremely uncommon, but then you have to take that and try and analyze that in the totality of the data.

And so the term signal is used

differently by different people, but I think that it's best used most accurately to simply state when there may be a concern, when there's a red glad raised as opposed to establishing a definitive and quantitative and comparable risk based on these reports.

The limits are, of course, it's a voluntary system. So there is under reporting. So while it potentially encompasses the universe of drug use, it really doesn't. It can't provide rates for rare events, and as I mentioned, looking at specific drugs, specific events, you can't generate comparative data.

And causality assessment is most difficult in these cases because the amount of data generally provided is not nearly as rigorous as a bedside clinician would want in assessing.

This just goes through the issue of causality and limitations in the case reports and the quality of the data that we frequently get.

So to get to the AERS database, there is a review in the background document that discusses

an analysis by ODS of 16 cases that were temporally associated with acute liver failure in the United States, in addition to international cases as well.

And the limitations of looking at individual case reports, again, are outlined here. There is the inherent subjectivity, and I don't use that in a pejorative sense, but in reality at the bedside for an individual patient unrelated to post marketing reports or clinical trials, clinicians do need to use their clinical skills, and that may be in a sense a synonym for subjective in assessing causality.

And there's been a lot reported. In the literature there are articles on the subject. There are instruments, causality assessment measurements trying to get at this.

And there was recently a meeting on hepatotoxicity actually in this city last month, and the issue of causality assessment is a prominent one. It's a concern, and it is an issue.

The analysis of these reports was done by ODS. I reviewed them myself, and again, there

was so much difficulty in assessing the relationship between drug and even that, in addition, we asked two external expert hepatologists to give us their views on these particular cases and on the panel here today.

My conclusions from looking at these cases are that there are, indeed, cases of probable leflunomide induced acute liver failure. So as a signal, using that term as I discussed, there is a signal. Events have occurred.

There are additional cases that vary from possible to unlikely in this database. There is confounding, meaning other possible, probable, likely, all of the above factors in the vast majority of these cases.

There was no consistent pattern across these cases that would suggest that it is, indeed, this drug that connects these cases one to another both in terms of clinical presentation, as well as the biochemical pattern of liver function test abnormalities.

And this doesn't mean that acute liver

failure cases haven't occurred truly related to the drug, but looking at a case series in a sense, this is distinctly unlike other series of hepatotoxins that the agency has reviewed and dealt with, such as troglitazone and bromfenac.

So the question for us is: do these cases represent the tip of an unreported iceberg or are they truly exceedingly rare events? And how can we quantitate the risk? Is the overall risk-benefit ratio for patients changed by these reports?

And ultimately we need to look at this issue in the context of other therapies.

The goal of the rest of my presentation is going to be an assessment of all of the available databases that I could find, that I could bring some evidence to bear on this issue, and to try and give you basically an evidence based assessment of what toxicity in a sense the highest estimate that we could find being very conservative.

And I will go through seven databases, first the clinical trials database, both the premarketing as well as post marketing studies.

Then separately I'll briefly discuss post marketing studies that were combination therapy. These are separated out really because the potential effect of combined therapy could impact the analysis.

In reality, the data isn't a lot differently, but they were assessed separately and will be presented that way.

There was a cohort study that was presented by the sponsor to the agency over the past six months, a cohort analysis, a second cohort analysis, and publication at the most recent American College of Rheumatology meetings in October of 2002 by the National Data Bank for Rheumatic Diseases.

And I've had personal communication with the author of that abstract, which is now in manuscript. It isn't published -- in an attempt to basically call that database as well for possible serious events.

There was a recent publication in the Annals of Internal Medicine in December of 2002 by the U.S. Acute Liver Failure Study Group, and after

reading that, I contacted the primary authors to see whether, again, within that database there were cases of acute liver failure associated with leflunomide.

And finally, the data mining analysis that I will go through.

Before I go into these cases, I want to try and keep the air as clear as possible on what I'll be referring to as serious hepatotoxicity. There is no one definition, and these are various possibilities.

To the extent possible, I will be using either hepatocellular necrosis associated with clinical jaundice, which has been termed Hy's Rule in the name of Hy Zimmerman who coined it years ago as a clinical pearl, and he's unfortunately now deceased, and by that definition if at bedside you have a patient who is presenting clinically and biochemically with hepatocellular necrosis and is clinically jaundiced, the mortality rate in his experience, and other authors have reproduced that experience, has at least a ten percent mortality

rate, and this has varied upward from ten percent depending on the particular etiology of acute liver failure.

Hospitalization for hepatocellular necrosis, which generally actually would be a less severe event than that, but that intuitively has a basis in definition for a series of hepatotoxicity.

Obviously acute liver failure and death. These are so rare that, you know, in looking at the realistic databases that we have, we can't rely on those events because studies that we could even conceive of would not really give us the power to identify those cases in controlled trials.

First I'll go to the clinical trials database. These were 17 controlled clinical trials between 1989 and 2002. Se requested the sponsor do a pooled analysis of these studies to maximize our power to see potentially meaningful differences among study groups. Kaplan-Meier, as well as an analysis of rates per 100 patient years was provided the sponsor.

The background document actually gives

an exhaustive presentation of all the various serious adverse events that have been discussed in the past in association with leflunomide. But for current purposes, I am going to be looking at clinically serious events using various definitions that I've presented.

This is just to give you an idea of the exposure. Ultimately power is the bottom line when you're looking for identifying rare events, and so I'll just briefly discuss what the exposure was so that we can get a sense of what the power would be here. to identify events of varying rarity.

There were about 1,700 patients exposed to leflunomide, 700 to methotrexate, 130 to sulfasalazine, 300 to placebo, and as you can see, if you look at as and 24 months, you do have fair numbers of patients if you're looking at events in the rate of one out of 100 or so and wanting to exclude the possibility of those occurring, and of course, these other groups are way too small to use going out further than a few months.

This is a Kaplan-Meier curve for ALT or

AST greater than three times normal. I apologize for the difficulty in reading it, but this is methotrexate in red. Leflunomide is in white, and the other two are, of course, the sulfasalazine and placebo, and remember the sulfasalazine and placebo in a sense end their exposure someplace down here.

There was a post hoc p value associated with this difference, but I do want to point out actually more so in the negative than in the positive this slide in that as has been referenced earlier, folate supplementation will decrease the incidence of transaminase elevations with methotrexate. So this curve really reflects what was seen in the clinical trials. If this was a curve that only looked at patient supplemented, this difference probably wouldn't be here.

But they do represent the data as they were done in the artificial setting of clinical trials.

This is an analysis of higher levels of transaminase elevation of ten times the upper limits of normal.

This is in the ballpark of what one would expect to see in what we'd call hepatocellular necrosis as opposed to simply transaminitis. If transaminase elevations of this magnitude are seen that are based on hepatocellular injury as opposed to cholangitis or metastatic disease or other causes unrelated, it would give us a better metric than the three times upper limits of normal.

And as you can see, over time the rates end up being similar. There aren't a whole lot of events. One could, I think, over interpret this into a difference in hazard rates over time between the two, but I won't go into that. I think the data points are too few

This slide is, again, meant to point out the limitations of using transaminitis as definitive endpoint. We clearly use them in early studies of drugs in Phase 1, 2, and 3 trials, but it's not the endpoint, and certainly it's not what we're most interested in today. We're interested in serious events for patients.

These three studies are actually

referenced in the label. The label has, I would say, a fairly exhaustive analysis of the clinical trials database for liver function test abnormalities. It's meant really to highlight the limits rather than what these type of data can show us in that depending on what study you look at, methotrexate may look better or it may look worse, and of course, placebo itself is going to have a rate of transaminitis.

And so we have to remember that there are background rates if we're looking at simple numbers.

Ultimately we really need to look at causality, and that's what I'm going to attempt to do in the remainder of the discussion of this database as well as the others.

Again, just a reminder. The Hy's Rule, jaundice associated with hepatocellular injury. I asked the sponsor to provide us line listings and narratives for all patients who had elevations of ALT of any magnitude in conjunction with bilirubins over 1.5, the upper limits of normal. This is well

below what a Hy's Rule case would be, but I wanted to be sure that we didn't miss anything, simply something being on the borderline.

And in reviewing all of those cases, actually there was one case that didn't, in fact, cross the threshold ironically, but I do consider that to be a meaningful case of hepatotoxicity, and that on review of that case appeared to be a treatment related episode of a patient who was clinically ill, did visit a hospital based on their illness, although there was no jaundice associated with it.

Next, the post marketing studies. There were two of them. One involved a two arm study, 130 patients in each arm. For the first six months one arm was exposed to both leflunomide and methotrexate, and in the second six months the arm that was not exposed to leflunomide as then in an open label fashion exposed to leflunomide so that in total you have 260 patients that were exposed during its six months at a minimum to combination therapy.

In addition, there was quite a large

study, 4,002, that looked at almost 1,000 patients for at least six months and patients who did not respond to an initial period of leflunomide had sulfasalazine added. The sulfasalazine ended up being quite a small population, and again, I'm using this really as a database to try and cull any cases of significant hepatitis.

Out of these 1,200 subjects, there were no cases of hepatocellular jaundice.

So in summary, reviewing all clinical trials, ALT elevation is not uncommon, in the range of two to four percent. ALT elevations to a greater extent are under one percent, and out of the nearly 3,000 patients that have been looked at in the controlled clinical trial setting, there was one case of what I would call hepatocellular injury. Again, it's not one of the Hy's cases that carries substantial mortality, but it was certainly a clinically ill patient. Then there were no cases of acute liver failure.

Next I'll look at retrospective cohort studies, two that were provided by the sponsor for

our review and one that is based on an outside manuscript by Dr. Fred Wolfe, who is here today.

Briefly, this was a retrospective cohort study. It's a claims database with linkage to medical, pharmacy, and laboratory data, and this is a critical issues when you have cohort studies that are based on claims and coding, having access to medical information is critical to assessing the credibility of that database and potentially giving some information on causality.

There were 40,000 patients with RA in that database. Not all 40,000 obviously were on these therapies, but again, just to give you the scope of the power of this study to look at clinically relevant events, there were about 2,600 patients on leflunomide, almost 10,000 on methotrexate, and DMARDs. This definition is not mine. It was the sponsors of this study, but these drugs represented almost 15,000.

In terms of the strengths of this study, as I mentioned earlier, case validation was part of this, and all severe cases of hepatitis, and the

definition of severe case was based on codes, and I'm not going to go through all of them, but what were considered to be codes of severity, and this was a critical list, were evaluated, and there was 100 percent agreement between what the codes came in as and then what the study personnel who went out to validate that found.

Twenty percent of the more frequent, but less severe hepatic events were assessed, and there was 83 percent validation or correlation between the coding and the records review.

It's a large study, but, again, you can look at is the cup half full or half empty. Is it large enough to detect something that occurs one out of 10,000 or 50,000 times? Clearly not, but it does expand a database that we can use for safety assessment.

Weaknesses, again, validation we need to be clear is not the same as causality. So validation meant, yes, indeed, this patient did enter hospital, did have transaminase elevations of whatever the validation criteria were, but that

doesn't clarify necessarily whether it's a drug event relationship or not.

And of course, there's channeling bias in these type of studies, and it's hard to say whether you end up having a bias for one group versus another.

I'll mention at this point that the sponsor is going to be presenting some comparative data using these databases. My purposes today are really, again, to look at serious events, to see whether in as many databases in as large of a total population as possible do we see hospitalizations, do we see cases of hepatocellular injury with jaundice, do we see acute liver failure.

So my analysis, in a sense is, we could say, complementary of simply different than what the sponsor will be using these databases for.

And the results show that there was one patient on leflunomide and two patients on methotrexate that had hepatocellular necrosis, and this comes out to be a rate of .04 percent in leflunomide.

There were no cases of hepatocellular jaundice, and again, there were no cases of acute liver failure.

Data on hospitalization is not available in this study. The next three databases will offer that.

This was, in a sense, two cohort studies that were looked at separately and the results from each separately are available in the background packets, and then they were looked at in combinations as well by the sponsor.

The databases were standardized claims data. Different managed care organizations. As you might expect, the Medicare database resulted in Protocare being a less well population, although the trends that the sponsor will show are similar regardless of which study, and I'll be looking at both in combination for my purposes.

There was as large database to sample 130,000 RA patients, 42,000 of whom were on a therapy, 2,800 on leflunomide, 15,000 on methotrexate, mean follow-up was well over a year.

So, again, in terms of power, it adds substantially to the database that I have tried to accumulate in my analysis.

The weaknesses similar to any cohort study, channeling bias, issue of causality assessment, there was not the ability to validate these cases as there was in the Aetna cohort study. It was simply the nature of this study.

The events that were included in this analysis requiring hospitalization related to these codes. An expanded analysis using these same codes but not requiring hospitalization was performed as a separate analysis by the sponsor.

There were no cases in either of these two databases of hospitalization for any hepatic event in leflunomide. The methotrexate group did have several. I think for our purposes, from my discussion, it's really what we're looking at again, the zero numerator for this particular severe definition in that size database.

This is the hepatic events not requiring hospitalization endpoints. This is the secondary

analysis, and this is a less precise and noisier type of analysis than hospitalization, but looking at it this way, there didn't appear to be a difference between the two drugs.

So in conclusion from these two studies, I'll say there were no cases of leflunomide related serious hepatitis defined by hospitalization for an hepatic event, and that included hepatocellular necrosis, which would be expected to include the universe of drug toxicity.

And the risks appear to be similar to the extent that a study with these limitations can tell us.

Next, the National Data Bank for Rheumatic Diseases. This is a nonprofit research organization, and this is a longitudinal patient reported surveillance program that actually started in 1998. Patients were recruited both from rheumatology practices around the country, as well as from a registry that was established by the sponsor Aventis.

Adverse events were collected from

patients by mail surveys every six months, and in order to be considered a participant, at least one semiannual questionnaire needed to be sent in, and you can see from the numbers of responses that on average it appeared the patients were in the ballpark of a year and a half on therapy during this period.

Hospitalizations and deaths were assessed through physician records as well as death certificates, although the initial ascertainment of a toxicity came from the patient surveys.

The strengths again are the size of a data bank like this and that the serious events were validated. Weaknesses, the same as you get from a data bank or a cohort study.

The results were for hospitalization rate for ICD-9 related liver codes was similar between the two groups. Now, if I were to prospectively define a study, I would really choose codes that are going to be more specific to drug induced hepatotoxicity. The ICD-9 liver related codes will, in a sense, by definition include some

events that are not going to be related to drug induced hepatotoxicity, and so it will give you a noisier estimate.

There will be inclusion of a lot of events that aren't really going to tell us anything about the drugs in question. Again, this database for my purposes was more importantly aimed at looking for cases of serious toxicity.

There was one patient who was hospitalized on treatment with leflunomide. That patient was neutropenic as well, was febrile. The ALT elevation was in the range of 500, and certainly this could have been the hepatopathy or the transaminase elevations may well have been associated simply with the underlying septic process, but for our purposes I'd like to be as cautious as possible, would include this as a case of hospitalization for hepatocellular necrosis.

There were no cases of hepatocellular jaundice or acute liver failure.

Again, this is a database with over 5,000 people, and that one case.

And finally, the publication in the Annals of Internal Medicine. The results of a prospective study of acute liver failure at 17 tertiary care centers in the United States in the Annals of Internal Medicine. This was a 41 month experience with a consortium, 17 liver transplant centers. It did cover the first 30 months of leflunomide marketing, and personal communication with Dr. William Lee, his estimation is somewhere between 25 and 40 percent of the transplant capability in this country is represented at these centers.

Now, I want to make it clear I don't want to misrepresent this. Number one, that was his estimation, and the other issue is that this is not the universe of serious hepatotoxicity or even death. It would represent whatever proportion of cases that are referred for evaluation for transplant, but it is, in a sense, a quantifiable percentage of the U.S. population that was included in this experience.

There were 308 cases or I should say

patients that were admitted to these centers with acute liver failure, and actually for purposes of the publication and for public health, the aspect of this particular experience that has received the most attention and appropriately so is that 40 percent of the cases actually were associated with the use of acetaminophen, and it highlighted the relevance of acetaminophen in the burden of acute liver failure in this country.

Thirteen percent of the cases were drug related, but other than acetaminophen -- and not surprisingly these were over represented by drugs we are aware which are no longer marketed. Four of the cases were bromfanac, four were troglitazone, and five were INH.

Again, this information, in fact, I don't believe is in the publication, but I've spoken with Dr. Lee and asked him what the breakdown was for these other non-acetaminophen cases.

There was one case of acute liver failure in that database that was associated with leflunomide, and interestingly that case was

captured by the FDA's AERS database, and it was the one case unrelated to overdose that in my own review I felt had probably the least potential for confounding or confusing, was a drug related acute liver failure.

My conclusion from this study is that while certainly there is under reporting, the extent of under reporting associated with acute liver failure, which is a very striking clinical presentation, may well be lower than that quoted in general for under reporting of adverse events, and in the literature you hear upwards of 90 percent under reporting. I think this experience suggests that may not be the case when talking about acute liver failure.

So my conclusion from an analysis of the hepatotoxicity and available databases is that in clinical trials, ALT elevations are as labeled present. They're consistent in both leflunomide and methotrexate use, and of course, in placebo groups this is not the ultimate endpoint of clinical importance for patients and doctors.

Clinically significant liver injury defined by hospitalization was looked at, and it's really three databases here I should say rather than four because one of the databases didn't allow us to look at hospitalization as the endpoint.

And out of over 10,000 patients, there were two with hospitalization for hepatocellular necrosis. That gives us a calculated rate of .02 percent, and if we're looking at, again, to try and be conservative, what might we be dealing with, and a kind of rule of thumb, a rule of threes is if you divide that exposure by three, it's unlikely that we would be missing events more severe than what we've identified in a greater than one out of 2000 frequency.

In terms of hepatocellular jaundice or a Hy's case, there weren't any out of a database, and this is of the four trials. Three trials was 10,000 and then we'll put the acute liver failure or hepatocellular jaundice case back into the denominator here.

There were over 13,000 patients in these

multiple databases, and again, if we are going to assume being cautious that patient number 13,701 would have been someone who experienced hepatocellular jaundice, a numbers needed to treat maximum in a sense would be one out of 5,000, trying to draw from database rather than a modeling.

Now, if one were to assume that -- and it is an assumption. I'll readily admit that hepatocellular necrosis without jaundice isn't the same thing as hepatocellular injury with jaundice in synthetic dysfunction, but if we're going to assume that we've got a case here, which we don't, but if this were to be one out of 13,000 or, let's say, one out of 15,000 and the lower confidence interval rate would be one out of 5,000 for hepatocellular jaundice, I would not expect that a rate of more than one out of 50,000 patients would die associated with that hepatocellular injury.

Again, I don't want to say a caveat. Ascertainment in these databases is not the same as it would be in a clinical trial. It is a robust attempt to capture that kind of information, and it

leaves us with a dilemma. There are rare cases of hepatocellular injury associated with hospitalization in databases where we can draw some confidence of event rates. There are very few cases in the post marketing experience of that most hard, most serious, most rare endpoint of acute liver failure.

And the question for us is how to capture the risk of that rare event and, in addition, for clinicians how do we capture comparative rates for toxicities of similar import.

Obviously if we redirect patients from one therapy to another, in a sense they're buying the toxicity of the next therapy they're going to, and it's not the purpose of my presentation to say exactly what that toxicity is going to be, but clearly if you move patients, whether it's to other drug DMARDs or biologic DMARDs, there are toxicities associated with those agents as well that we have to put into the mix when making risk benefit assessments, as well as risk communication.

So for the patient who experiences acute

liver failure clearly there is no risk benefit analysis that's going to favor therapy. The issue for us is for prospective prescribers and patients. How do we interpret the magnitude of risk for very rare events, both the rare events as I define hospitalization and that we can estimate in clinical trials, as well as the uncontrolled databases and post marketing reports. How do we characterize these events for patients and physicians?

And my last analysis is going to be going back to the post marketing database in an attempt to look at that in a systematic way.

What is data mining? It's a system to allow computer analysis, and it could be of any database. For our purposes it's the AERS database that has millions of reports in an attempt to identify and quantitate signals for drug associated adverse events.

And I highlight signals here both in reference to my earlier comments about a signal was not a definitive statement of absolute risk, but a red flag for further evaluation and to highlight

again that there aren't absolute risks that we can quantitate out of a data mining analysis.

This is currently being evaluated in the Office of Biostatistics as a screening tool, and it does require further examination. There is a publication, along with several others that were sent to the committee. This was entitled "Use of Screening Algorithms in Computer Systems to Efficiently Signal Higher than Expected Combinations of Drugs and Events in the U.S. FDA's Spontaneous Report Database."

This is clearly a title that was written by someone in biostatistics.

(Laughter.)

DR. GOLDKIND: If I were titling thing article it would have been "Digesting the Data."

(Laughter.)

DR. GOLDKIND: Again, to remind everybody, there are strengths of the AERS database, and there are limitations. In deference to the time, I won't repeat the list.

Just to give us an idea of what is the

potential database that we're dealing with in the AERS post marketing system, and now this slide refers to leflunomide specifically, approximately two million prescriptions have been written since approval, and this represents between 250 and 300,000 patients. This number is a little bit older, the more recent data.

To remind everybody this is the universe of exposure, and 16 reports of possible acute liver failure were identified by ODS, U.S. based cases, and then 13 international cases that have been analyzed in the background document.

Why do we need to data mine? Well, to put into context these cases that you have probably been confused by reading the background document.

What do we take away from them? The attempt with data mining is to coherently organize and interpret a large database. How large is this database? Pretty darn large. There are over two million reports in Medwatch, and that is for 8,000 products, 7,000 preferred event terms, and if you'll conceptualize a two-by-two table of events on one

axis and drugs on another axis, there are 56 million potential combinations of a drug associated with a particular event.

There are 300,000 new reports that come in annually.

I'm going to just give you an example of what a data mining graphic display would look like, and of particular relevance to the issue at hand here. Dr. Szarfman kindly performed this analysis for us, looking at the term "hepatic failure." That was the event code used in the search, not mortality from hepatic failure, but hepatic failure.

And, of course, hepatic failure can be associated with many other terms, liver related terms. So the analysis can spread across, can be broken down, for the purpose of this analysis, which was hepatic failure, and only drugs that had at least three or three reports in this two million person database were going to be signaled.

And there is a color coding system that's used just to allow the human eye to graphically scan data, and on this slide, gray,

regardless of the shade, represents drugs that have been reported and at least if you see, there's going to be at lest three reports that have been reported, but looking at the ratio of reports for that drug related to hepatic failure and that drug's entire experience with adverse events in the context of the entire database did not signal as a higher rate than you would expect background if your null hypothesis was that all drugs would be associated to the same non-causal extent.

And this is actually only one page out of 17 in this particular analysis, and this analysis started at the earliest time point. So page 1 I don't recall, possibly going back to the 1960s or '70s would have been the very first drug to have three cases of hepatic failure, and I don't remember what page number this is, but we pick it up in 1997, and as you can see, troglitazone and bromfenac, which were marketed around this same time in the retrospective peak at what the experience reported in real time in 1997 and '98 were picked up as drugs that had a higher reporting experience than one

would expect it.

There are a lot of drugs in here in the sense that they can provide a negative controlled force for drugs that haven't been identified through other means as major hepatotoxins.

This is just a later page I did want to pick up. Leflunomide appears on the list. We all know, of course, there are more than three reports. We got the third report here in 1999, and these are cumulative total numbers in the system.

These were culled to exclude duplicate reports, but causality assessments are not part of this analysis. So these really are crude reports, and to the extent that causality is or is not assessed, it's equally across the database.

But leflunomide did not signal as a greater than expected signal for hepatic failure events.

The next analysis that Dr. Szarfman did was look at signals for hepatic failure, and this was meant, again, to -- I'm sorry. This slide is actually a summary of the previous analysis. I

didn't show all 20 pages, thankfully so, but in those 20 pages, there were signals for these commonly understood hepatotoxins.

The next analysis that Dr. Szarfman did for us was in relation specifically to rheumatoid arthritis, and I'm going to be showing this for several purposes. One is to highlight graphically the complexity of assessing post marketing serious and life threatening events, and another, it may provide some insight into the AERS reports of serious hepatic events for leflunomide.

These therapies that are used in RA were analyzed in addition to some control drugs, again, which have been identified based on individual case reports and assessment through ODS as significant hepatotoxins.

Actually what we did in this analysis was to look not only at liver related events, which is our concern for today, but, again, to remind us of the context of multiple therapies and various toxicities being highlighted or toxicities of most concern for different drugs.

There are three different analyses that will follow in rapid succession. The first is liver related events. The next are opportunistic infections, and the third is lymphoma, and these are analyses of fatal events related to these systems.

Only drugs that are actually signaled as greater than you would expect show up in each slide. So you don't have every drug in every slide. As you can see, not every drug is here, and actually the rheumatoid arthritis therapies are under represented, which you would expect since we had positive controls, which are highlighted here, just to assess the sensitivity.

Leflunomide, again, did not signal in this system for fatal hepatic events. These are the various codes that come within the umbrella of hepatic events, and again, I don't want to go through each one. The purpose of this slide is to point out that those drugs that we have confidence are associated with hepatotoxicity were picked up in this system, and the leflunomide did not signal in any of these categories or for the umbrella of

hepatic events, fatal hepatic events.

The next is opportunistic infections, and there was a lot more discussion of this yesterday, and you see there's a different fingerprint in the sense for drugs, not surprisingly.

A couple of points I want to make on this slide. One is while aspergillosis was picked up, there were seven cases of leflunomide. There's a stronger signal, again, as pre-marketing, post marketing would have expected across the biologics.

The important other thing to mention here is when you have a drug that's used to treat various diseases, you have to take that into account when trying to analyze these data, and data mining is a computer system, and this one at this point in time doesn't take that into account. So you can't really look at methotrexate as an RA therapy in the context of this database.

Many of these cases are probably related to methotrexate and used as an oncolytic agent at higher doses with more immune suppression in

conjunction with other immunosuppressive agents, and also, INH and Rifampin you would expect there would be more reports since those drugs are used to treat the disease.

So this really simply highlights that this is not a tool that allows us to be mindless about analysis. You obviously have to bring some knowledge to this database and query it. If you have a signal, you have to say, "Okay. What might that mean?" And then it becomes an exercise really of case study of the individual reports.

And finally, lymphoma. There are signals in this database for fatal outcomes associated with lymphoma for taniceptin and infliximab. Again, methotrexate is a therapy. It's for lymphoma. So not at all surprising, that would signal the highest, and you take that into account when you figure out what these data may mean. Likewise prednisone is used in oncology as well.

So my conclusions from these data mining analyses are that it is a tool that's currently under evaluation in most marketing safety

assessment. These signals do require interpretation and validation based on review of reports. Certainly false positive signals in a sense will be identified, and the identification of them as false positives really only follows a more detailed, thorough analysis of the case reports themselves.

False negatives, in the analyses that Dr. Szarfman has done, I really don't believe there have been any in the analysis she's done, but this really is still undergoing assessment.

I feel that it does graphically highlight the complexity. it wasn't meant to confound and confuse, but only to share with you how difficult it is to put post marketing reports into the context of drug causality, as well as the context of therapies that are available.

And it does, I think, convincingly identify how each drug is going to have its own unique toxicity profile and, again, it's a multidimensional analysis of what drugs are appropriate for marketing, what drugs are appropriate for what patients.

Now, leflunomide was not identified above a threshold for a greater than would be expected rate in these analyses, while other drugs that we generally have a consensus are at a high level of serious hepatotoxicity showed up.

This does not at all mean that acute liver failure has not occurred or cannot occur with leflunomide. It does suggest that the pattern of reported hepatic failure events for leflunomide is different than that for other drugs with known and clear hepatotoxicity, such as troglitazone, trovafloxacin, valproate, flutamide, isoniazid, or bromfenac.

So in summary, my overall conclusions regarding hepatotoxicity and leflunomide are that the biochemically defined hepatotoxicity of ALT elevations greater than three times normal are not uncommon; that serious drug induced hepatotoxicity defined by hospitalization in databases where we really can have a data driven rate to calculate are rare in these three data bases that we looked at, .02 percent.

Acute liver failure and death have been reported in the post marketing experience. We cannot establish a rate based on those isolated reports, and cases of hepatocellular jaundice did not occur in these large databases. So we can't really quantitate what the rate would be, but looking at the 13,000 patients that were analyzed in these databases, in a sense we can say what the rate is not likely to be, and again, as I had mentioned earlier, if we were able to assume the patient, 13,701 were to have hepatocellular jaundice, it's unlikely that the frequency of that event in association with this drug would be more than one out of 5,000, and if we're going to take a ten percent mortality for hepatocellular jaundice as a rule of thumb, we would estimate that the rate of death due to acute liver failure with leflunomide use would not be more than one out of 50,000.

When looking across drugs used to treat rheumatoid arthritis, life threatening events are associated with all therapies. We didn't even touch on NSAIDs today, but probably the audience as well

as the panel is well enough aware of the potential risks of NSAID.

Obviously DMARD drugs and biologics, and this was obviously discussed in more detail yesterday, clearly all have their potential safety concerns and risks that are being weighed in in patients on therapy.

It's important for us and particularly uniquely as the FDA, as the regulatory agency involved with risk communication, to characterize and communicate these rare but life threatening events as coherently as possible for optimal use of these drugs.

Thank you.

CHAIRPERSON ABRAMSON: Thank you very much for a very comprehensive and, in fact, scholarly presentation.

I'm going to move on to the next presentation by Aventis Pharmaceuticals and Dr. Rozycki will present.

DR. ROZYCKI: Good afternoon, ladies and gentlemen, and once again, on behalf of Aventis, I'd

like to thank you for the opportunity to be here today to discuss the safety issues that have arisen with regard to Arava.

Our presentation this afternoon is going to focus on the benefit-risk profile of leflunomide or Arava, and if you look at the different parts of the benefit-risk equation, on the benefit side we feel that leflunomide is an effective and unique treatment for rheumatoid arthritis. It has a unique mechanism of action. It's already indicated to treat signs and symptoms and to retard radiographic -- I think that should be to retard structural damage.

And as was discussed earlier today, the possibility of adding to the indication for improvement in physical function, and as Amye Leong so eloquently described earlier today, we feel very strongly that it provides a critical therapeutic option for patients with rheumatoid arthritis who don't otherwise have that many options.

On the safety side of the equation, the leflunomide safety profile we feel is well

establishes between what is in the current labeling and what is in ongoing discussions with the FDA for labeling, and again, this is not to say that it is without adverse events or serious adverse events even, but that it is an established safety profile.

So taken together, as we will discuss through the course of the afternoon, we feel the benefit-risk profile for leflunomide is comparable to that of other DMARDs and justifies its continued use in the treatment of rheumatoid arthritis.

Just as an overview of our presentation, Dr. William Holden of Aventis' Epidemiology Department will provide an overview of the AE rates for leflunomide compared with other treatments. As Dr. Goldkind explained a short time ago, there will be some overlap between the data sources that Dr. Holden will discuss and that Dr. Goldkind discussed previously.

But, again, Dr. Holden's emphasis will be on a more broad based view of the epidemiology and benefit-risk of leflunomide.

Following Dr. Holden, Dr. Vibeke Strand

will provide a rheumatologist's view of the overall benefit-risk of leflunomide, and then we'll wrap up.

So if I could introduce Dr. Holden.

DR. HOLDEN: Thank you, Dr. Rozycki.

Good afternoon, Mr. Chairman, ladies and gentlemen of the committee. My name is Billy Holden from the Aventis Global Epidemiology Department, and I'd like to spend this part of the presentation discussing the ongoing activities in pharmacovigilance and epidemiology that we've taken with regard to leflunomide.

I'd like to first discuss a pooled analysis of the Phase 2 and Phase 3 clinical trial data, then move on to a brief discussion of spontaneous reports and post marketing data, and from there discuss and spend the bulk of the presentation discussing two large epidemiologic studies that we did after we analyzed the early post marketing data.

The pooled analysis relied on data from the Phase 2 and Phase 3 pivotal clinical trials, some of which were described earlier today. There

were five Phase 2 trials, which included 550 patients, mostly taking leflunomide.

There were five Phase 3 trials which included over 2,300 patients, half of whom were taking leflunomide. The data from these patients were combined into one data set, and cumulative rates per hundred person years were calculated for different events.

So there were a total of over 2,800 patients in the combined analysis accounting for about 4,400 person-years of exposure.

The first set of slides compares leflunomide to methotrexate on Labbe (phonetic) scatter plots or line of identity graphs. These graphs are interpreted by finding data points to the left or above the line, which would indicate higher rates for methotrexate, and conversely points to the right or below the line, which would indicate higher rates for leflunomide.

And what we can see here after six months is that leflunomide has slightly higher cumulative rates of infection, pulmonary

hypertension, skin, and hepatic serious adverse events when compared to methotrexate.

Methotrexate treated patients had slightly higher rates of malignancy and cardiovascular and thromboembolic events.

After 12 months the cumulative rates follow the same pattern, although now hepatic adverse event rates are equal, and after 24 months the patterns persisted, although differences in pulmonary and infection are actually quite small.

We then looked at hepatic events in more detail, and here hepatic refers to all of the events captured by a series of predetermined COSTART codes and includes both serious and non-serious events. And by serious I mean the regulatory definition, which includes events that resulted in hospitalization, disability and death.

The transaminase elevation data actually came from a separate set of laboratory results, but some of these results could have been captured in the hepatic adverse event code on the top if the treating physician reported them as adverse events.

And what we can see here is that methotrexate clearly has much higher rates of all hepatic events and transaminase elevations, including three times, five times, and ten time the upper limit of normal.

At 12 months we see that this pattern persisted, and again at 24 months we see that this pattern persisted.

We repeated this entire analysis, this time comparing leflunomide to sulfasalazine, and at six months we can see clearly that all of these serious adverse events that were reported, with the exception of cutaneous, were more common amongst the sulfasalazine patients.

At 12 months, only cutaneous and infection are higher amongst the leflunomide, and at 24 months, cutaneous and infection continue to be higher in the leflunomide group and the rate of cardiovascular and thromboembolic events is very slightly higher amongst leflunomide users as well.

When we looked at hepatic adverse events, overall events are more common among

leflunomide patients. Rates of transaminase elevations are clearly higher among the sulfasalazine users at six month.

And at 12 months both hepatic events and enzyme elevations are more common among sulfasalazine users, and at 24 months, hepatic events and mild enzyme elevations are more common amongst the leflunomide patients.

So what can we conclude from this analysis?

First, compared to methotrexate, leflunomide had comparable rates of serious hepatic adverse events, possibly high rates of hypertension and cutaneous events. Leflunomide users also had lower rates of all hepatic events and transaminase elevations through 24 months.

Compared to sulfasalazine, leflunomide had fewer serious adverse events except for infection and cutaneous events. Transaminase elevations were more common amongst sulfasalazine users, although all hepatic events were slightly more common amongst leflunomide patients.

Several signals were generated from these data that relied on all of the available safety data from Phase 2 and Phase 3 clinical trials, but overall there was no clear demonstration of an increase in risk for leflunomide.

After the drug was launched in the fall of 1998, we started our pharmacovigilance activities, which included Phase 4 clinical trials, epidemiologic studies, the development and implementation of risk management programs and intensive reviews of spontaneous reports and other post marketing data, all performed by a dedicated safety staff.

I'll briefly review some of these post marketing data.

Everyone here is familiar with the limitations and biases inherent in spontaneous reporting. Just to mention a few, the adverse event that's reported may not be related to the drug. This caveat, in fact, appears on the Medwatch reporting forms and may be related more to the underlying disease.

Reporting rates themselves are not measures of incidents or occurrence. They are measure of reporting intensity, and the many factors that affect the actual reporting of spontaneous events, such as the severity of the event at the time the product has been on the market and the health care professional inclination to actually file a report, all contribute either to under reporting in most cases or occasionally perhaps even to over reporting.

We take spontaneous reports very seriously, and we use them for several activities, including the prioritization of safety reviews. These events are reviewed in more detail, and some of them are singled out for telephone and the questionnaire follow-up.

Spontaneous reports aid in the identification of signals which we use in further studies, and they facilitate discussion with regulatory agencies around the world and focus endpoints for epidemiologic studies.

What we can see here is the U.S. and

rest of the world exposure to leflunomide, and we use these data for denominators in calculating reporting rates, and basically what we see here for both the U.S. and globally is that there's a steady increase over time in the exposure to leflunomide, and these data can be interpreted in one of two ways. Either more patients are being exposed to the leflunomide or more patients are using the drug for longer periods.

These data are through September 2002, and there are approximately 405,000 person-years of exposure. Through December 2002, although not represented here, there are about 450,000 person-years of exposure.

Here are the reporting rates for acute hepatic failure, and what we can see here, first of all, is that relative to infliximab and etanercept the rates are comparable, and we looked at these two biologic DMARDs because they were launched at approximately the same time as leflunomide.

What we can also see here, looking at the yellow squares on the bottom of the graph on the

left, which represent the reporting rate for methotrexate, this underscores one of the hazards of using spontaneous reports for reporting rates, which is that even though we know that this drug causes hepatic events, because it's widely prescribed and because it has been on the market for 50 years and prescribing physicians are familiar with its toxicity profile, very few events are actually reported.

In epidemiology this is known as a secular trend problem. Specifically in pharmacoepidemiology, this is an extreme example of the Weber effect, which states that spontaneous reports diminish considerably after the first two years a product has been on the market.

Also, the hepatotoxicity of methotrexate may be more chronic than acute, and this would contribute to its under reporting, although there are, of course, cases of acute liver failure in RA patients receiving methotrexate.

We can also see in the box on the right the cumulative reporting rates which, again, confirm

that the leflunomide and infliximab and etanercept have approximately equal reporting rates for hepatic failure.

Another way of looking at these data is to look at the actual number of cases reported, and what we can see here is that there are, in fact, cases reported for methotrexate; in fact, more so than the other comparator drugs.

The point here is twofold. First, acute hepatic failure is reported with all of the DMARDs; and, second, we should view these data with caution, especially when reporting rates are calculated.

Another source of post marketing data is the United State Network for Organ Sharing, which is an organization that oversees transplants in the United States and has been collecting data on transplants since 1986. It has a large database and has been collecting data on organ transplants since that time.

We looked at liver transplants from 1998 through July of 2002, and when we looked at what UNOS calls the etiology of the liver transplant, we

found 15 transplants listing methotrexate toxicity. In that same time period we found none for leflunomide.

However, we are aware of two cases of liver transplant associated with leflunomide. One is a recent case from Italy. So it would not have been captured in this database.

The other occurred in the fall of 2002 in the U.S., but because these data and prior, it was not captured here. This case, however, is very confounded, and it's not clear that leflunomide in any event would have been listed as the etiology.

And later I'll show some examples of some typically confounded cases, which are the norm in our post marketing experience with this product.

So based on our analysis of spontaneous report and other post marketing data, as well as on the signals generated from clinical trial data, we decided to do an epidemiologic study to quantify the risks involved with using leflunomide.

The first study we did was a retrospective cohort study using Aetna claims data.

Aetna is a managed care company in the United States, which covers six and a half million lives. It has a large database with links between medical, pharmacy, and lab data. It captures all in-patient and hospital diagnosis claims, as well as all dispense prescriptions for its members.

We chose the Aetna database for two reasons. First, it had by far the largest number of leflunomide users, well over 5,000, more than any other database that we examined when we initiated the study in early 2001. And we examined all of the publicly available databases in the United States and in Europe.

For example, the database with the second highest number of leflunomide users, United Health Care, had only about 1,900 leflunomide patients at that time. The GPRD in the U.K. only had 200 users.

The second reason that we chose Aetna was because it allowed access to source medical records, which we needed for case validation purposes.

The time of follow-up in this study was September 1998 through December 2000, and rheumatoid arthritis and diagnoses were identified through ICD-9-CM codes.

The cohort itself was defined as all patients diagnosed with rheumatoid arthritis who had received a DMARD. Patients had to be 18 years of age or older. The date of first DMARD rescription had to be after September 1st, 1998, and we excluded from the cohort patients who had experienced any of the hepatic events of interest in the three months prior to the start of the cohort.

The primary endpoints in the study were hepatic events. We looked at hepatic necrosis, hepatic coma, noninfectious hepatitis, hepatocellular jaundice, cirrhosis, elevated enzymes, and some nonspecific liver disease codes.

The secondary endpoints in the study included serious cutaneous disease, hypertension and respiratory infection, hematologic disease, and pancreatitis.

Exposure was measures through dispensed

prescription data, and we defined several exposure groups in this study, including leflunomide, methotrexate, and DMARD monotherapy. The DMARD group includes biologic DMARDS, etanercept and infliximab, as well as sulfasalazine, hydroxychloraquin, penicillamine, gold, minocycline, cyclophosphamide, and cyclosporin.

We also looked at three combination therapy groups: leflunomide plus methotrexate, leflunomide plus other DMARDs, and methotrexate plus other DMARDs.

Covariates that we used in the analysis included age, gender, and comorbidities, which we measured using a modified Charleston index, as well as the actual numbers of comorbidities.

And the analysis included a simple description of a cohort in terms of age, gender and person-time, and we used Poisson regression to estimate incidence rates.

And before I present the results, I want to talk about the limitations of the study. We did not have indicators of disease severity. We had no

direct measures of HAQ scores or joint counts, things of that nature. And, in fact, we had limited clinical detail.

We did not have data on the history of rheumatoid arthritis, prior treatments or hospitalization, and we did not have data on over-the-counter medication use, and of course, we had no data on actual adherence to therapy.

We were not able to pull out the biologic DMARDs from the others, not because we didn't want to. We did, but because we did not have direct access to the raw data due to privacy concerns and had to work through an intermediary who passed all of our analytic requests to Aetna.

We identified in the database 40,594 RA patients. The crude prevalence in the database was 0.6 percent. Three quarters were women. Most were in the age range of 51 to 64. About 80 percent of these patients were on monotherapy or two drug combination therapy.

And this is not different from what one would see in a typical rheumatology practice. So

these results are both generalizable and characteristic of other data sets.

We had a total of over 83,000 person-years of follow-up making this the largest rheumatoid arthritis cohort study ever performed. DMARDs alone or in combination accounted for 72,000 person-years of follow-up, and leflunomide alone and in combination accounted for over 11,000 person-years of follow-up.

The exposure groups themselves were comparable in terms of age, gender, and mean exposure times. The mean exposure time of patients on leflunomide in this study was about 18 months, similar to the other exposures, a little less than the DMARD group, which had about a two year mean. And this year and a half mean exposure time is in accord with published data and presented data on exposure times to leflunomide.

In terms of comorbidities, again measured at baseline and at the time of the event, the rates were comparable between leflunomide, methotrexate, and DMARD.

Because our primary endpoint focus was hepatic events, we validated a 20 percent sample of these claims used in the analysis, and we found 100 percent agreement between the data in the medical records and the claims that were submitted for hepatic necrosis diagnoses, and over 80 percent for all of the diagnoses.

The validation process is described here. Aetna requested the necessary medical and other records, including labs offering a financial incentive to respond. Data were de-identified, and a trained clinical assessor reviewed them and entered required data onto forms developed by the FDA, Pharma, and the American Association for the Study of Liver Diseases, which I will show briefly.

I'm sorry if this is hard to read. Basically the information captured here includes history, prior hepatic disease, drugs used, lab tests, and other results, and on the second page there's data on comorbidities, as well as occupational and environmental exposures.

The validation effort was labor

intensive and very time consuming, and such efforts were critical to the validity of a study becoming increasingly difficult due to HIPAA and other patient privacy legislation.

We can see here the overall cohort rates for the various endpoints of interest, and what it shows is what we know about the natural history of rheumatoid arthritis. In other words, this is a relatively sick population, one that carries with it an excessive burden of illness.

And this, by the way, is one of the great challenges of doing epidemiologic studies in rheumatoid arthritis. It is extremely difficult to distinguish between the intrinsic effects of RA and the effects of the medicines that are used to treat it.

Any endpoint experience was about 140 per thousand person-years of exposure, and here any endpoint refers to the limited number of endpoints that we included in the study. So this rate underestimates what's happening to this population.

For hepatic events, there are about

eight per thousand person-years. They were relatively high rates of hypertension and respiratory in this cohort.

When we focus on the cumulative hepatic rates among the different treatments, and these rates represent a mix of chronic and acute liver effects, what we see is that there's no difference between any of the exposure groups, and this includes the monotherapies, as well as the two drug combination therapies.

When we focus on more severe hepatic events, this slide shows very clearly that the rates for hepatic necrosis, hepatocellular jaundice, cirrhosis and noninfectious hepatitis are virtually equal across the board.

And again, when we further drill down to hepatic necrosis where we had 100 percent agreement on the validation form, we again see, despite the low numbers, that there's no difference between the three main exposure groups.

Although time doesn't allow me presenting them as pattern of results in which we

saw for leflunomide users compared to the rates in other DMARD users, we saw a comparability of rates for every endpoint that we examined, including severe cutaneous disease, hypertension, respiratory, hematologic, and pancreatic events.

Again, this was the largest rheumatoid arthritis cohort study ever performed. It was performed in a closed system in which all members are known, all demographics are known, all dispensed DMARDs are captured, one in which in-patient and out-patient diagnosis claims are captured, and one in which we could validate certain outcomes.

The design of the study allowed us to follow changing medication patterns in patients and measured directly the strength of the association between the drug exposure and different endpoints.

These facts, of course, do not prevent channeling bias, the phenomenon that occurs when patients with different levels of disease severity are preferentially prescribed one drug over another. Although it's difficult to hypothesize about theoretical biases in a study, in this case it may

be that patients with more severe RA were , in fact, channeled to leflunomide. Leflunomide was the first new DMARD in a decade, and no DMARD works consistently for the long period of time that the disease persists.

It's not unreasonable to assume then that many RA patients perhaps sicker than the rest were put on leflunomide. The channeling effect would result in an exposure group with more severe RA than the others, and bias this study against leflunomide.

But the bottom line and the take-home message from this study is that the rate of hepatic and other endpoints that we saw in the leflunomide exposure group were comparable to the rates in the other DMARD exposure groups.

Aventis wanted to replicate the study. We asked Professor Sammy Suissa of McGill University in Montreal to do a second study for us and to do it independently. He has given me permission to present the results of his study, although he is here himself to answer any questions about it.

The design of Professor Suissa's investigation is a nested case controlled study, which means it's a case controlled study performed in a predefined cohort of patients.

The cohort itself came from a combination of two very large databases. Again, these are claims databases from U.S. managed care companies covering about 26 million lives in total.

The time of follow-up was September 1998 through the end of December 2001, and again, rheumatoid arthritis and diagnoses were determined through ICD-9-CM codes.

The cohort was defined similarly to the way we defined it in the Aetna study. Patients have to have an RA diagnosis. Patients have to have a prescription for DMARD after September 1st, 1998. Patients had to be 18 years of age or older at the time of entry into the cohort. Patients needed three months eligibility prior to entering the cohort, and again, patients who experience any of the endpoints of interest in the three months prior to entry were excluded from the cohort.

The cases or endpoints in this study were of two types. The first required hospitalization, and these included hepatic, hematologic, cutaneous, lymphoma, infection, pancreatitis, and pneumonitis events.

The second type of case did not require hospitalization. Cases were both out-patient as well as hospitalized, and they included lymphoma and opportunistic infection.

Controls were matched ten to 100 on the date of the cohort entry, and of course, they had to be at risk for the event on the day of the case event.

Exposure, again, was identified from dispensed prescription data.

Professor Suissa defined several exposure groups in this study, including methotrexate monotherapy, which was used as the reference, leflunomide monotherapy, and in combination with other DMARDs, which include hydroxichloroquin, sulfasalazine, gold, minocycline, chlorambucil, penicillamine, cyclosporin, and

cyclophosphamide; a separate biologic DMARD group, including etanercept and infliximab, and in this study NSAID and Cox-2s and glucocortocoids were used as covariates in the analysis rather than as separate exposure groups.

Other covariates in the study included age, gender, the source of the data, comorbidities, and the non-use of DMARDs in the year prior to the event. The analysis itself relied on conditional logistic regression to estimate relative risks during the year prior to the indexed event.

The reference for the relative risk analysis is methotrexate, which by definition has a relative risk of one.

Professor Suissa also defined current use of leflunomide as a prescription within 90 days of the indexed event and past use of leflunomide was defined as any other use during the prior year.

Again, some of the limitations of this particular study, despite its size, certain diagnoses were very rare. Serious cutaneous events, there were only three: interstitial pneumonitis, 12

cases, and lymphoma, five cases.

There was no ability to validate the diagnoses in the study. These are proprietary databases, and they did not allow access to the source medical records.

The cohort itself included about 42,000 RA patients. The mean age was 49 in one database and 59 in the other. Again, about three quarters of the cohort were female and about 15 percent had used leflunomide at any time during follow-up.

There was a total of about 51,000 person-years of follow-up in this study. These are the total cohort event rates. They're on a different scale than the Aetna study. Again, these are hospitalized cases. So the rates would be smaller.

Any event experience was about 90 per 10,000 person-years, five per 10,000 for hepatic events; hematologic about 30; and infection about 42 per 10,000 person-years of exposure.

Now, let's focus on the serious hepatic events, that is, hepatic events that resulted in

hospitalization. Again, in this analysis, everything is relative to methotrexate monotherapy, which has a relative risk of one. While there were seven cases amongst the methotrexate monotherapy group and two cases amongst the leflunomide group, this resulted in an adjusted relative risk of 0.9, with a wide confidence interval.

The relative risk as adjusted for age, gender, the claims database from which the case arose, nonuse of DMARDs in the prior year, and the use of NSAIDs, Cox-2s, and glucocorticoids.

Two leflunomide events that occurred did occur in combination use, which didn't radically alter the relative risk. It went to 1.6 with an even wider confidence interval, and they both occurred in the past as defined by Professor Suissa, resulting in an elevated relative risk of 2.6, but with an even wider confidence interval.

Although not the main focus of this study, of no small interest here is the elevated risk that was seen for biological DMARDs of 5.4, with a confidence interval of 1.2 to 25.

The two leflunomide cases are presented here in narrative form. The first was in a 77 year old female who had received methotrexate for at least two years, and hydroxychloroquin for ten months prior to getting leflunomide therapy. She had received only a one month prescription for leflunomide nine months prior to being hospitalized. She received azathioprine two months prior to being hospitalized, and her hospital diagnosis was of acute and subacute necrosis, unspecified hepatitis, hepatic coma and respiratory abnormality.

The second case occurred in a 55 year old male who had received methotrexate therapy for at least six months prior to getting leflunomide. He had received leflunomide prescriptions for seven months, which ended ten months prior to hospitalization. He continued methotrexate therapy until two months prior to hospitalization, and he also had azathioprine therapy added four months prior to being hospitalized, which continued up to his hospitalization, and his hospital diagnosis was of abnormal liver tests and non-alcoholic cirrhosis.

And these point of these narratives is to demonstrate how remarkably confounded they are, and again, in that regard, similar to the spontaneous reports that we get.

Again, time doesn't allow me to present all of the data, but this pattern of no increase in risk was seen for the other endpoints in the study. What we saw, again, no increase in risk for all serious events, serious hepatic events, serious hematologic, pancreatic or opportunistic infection, septicemia events.

So to summary some of the results of the pharmacovigilance and epidemiology efforts that we've taken, the pooled analysis of the Phase 2 and Phase 3 clinical trials showed that the adverse rates of leflunomide were comparable to sulfasalazine and methotrexate.

Analysis of the post marketing surveillance data showed that the hepatic failure rate of leflunomide was comparable to other biologic DMARDs.

The Aetna cohort study showed that

hepatic and other event rates of leflunomide were comparable to rates of other DMARDs, and the nested case control study corroborated this by finding that there was no increase in risk of serious hepatic and other events in the leflunomide exposed group relative to other DMARD groups.

Now, in epidemiology we're trained to see the forest through the trees. We try to put things in context by getting a feel for the data, all of the data that are available and relevant to address an issue. The issue here is the safety of leflunomide relative to the other DMARDs.

The analyses presented here each have their strengths and weaknesses. Individually they provide incremental pieces to a larger puzzle. We are not claiming that leflunomide is without toxicity. What we are claiming, based on the analyses presented here, the forest, if you will, is that relative to the other DMARDs, leflunomide is just as safe.

Thank you.

Now I'd like to present Dr. Vibeke

Strand, who will talk about the benefit-risk profile of leflunomide.

CHAIRPERSON ABRAMSON: Excuse me. As Dr. Strand is coming to the podium, I just want to say that because we're running a bit late, we're going to work through the break. So anyone who wants to take a personal break during this time can feel free to do so.

DR. STRAND: So as you all get up to leave the room --

(Laughter.)

DR. STRAND: -- I will now try to give a perspective from a rheumatologist's point of view of the benefit-risk profile of this product.

I think we all know rheumatoid arthritis is a unique and severe disease to a heterogeneous population. We know that our patients have long-term deterioration in physical function and health related quality of life, but two year data is relevant even in the context of 20 or 30 years of disease because we haven't had two year data until the last several years, where we've now had five new

DMARDs introduced.

Current practice has clearly changed. Our aim is now to halt disease progression, and we certainly want to improve physical function and health related quality of life.

There's still a need for more therapies in rheumatoid arthritis despite the five new DMARDs or DMARTs, as Dr. Simon mentioned this morning. Not every one of them works in every patient. Not every patient responds to every therapy. As we've talked about several times, they have a long duration of disease with a long-term loss of function and loss of ability to work inside or outside the home.

There are few, if any, spontaneous remissions and few, if any, cures. I think what's most important is that tachyphylaxis develops with this disease to almost every therapy, and I think that was a very striking point that Dr. Fries pointed out to us this morning when he showed HAQ data with methotrexate therapy long term.

Leflunomide I think you have heard and discussed and decided even that it does have some

demonstrated efficacy. We know that it inhibits X-ray progression. It relieves the signs and symptoms of rheumatoid arthritis, and it also improves physical function and health related quality of life, but the point really is that it's comparable to methotrexate, our gold standard, and it's comparable to the biologic DMARDs or, shall we say, the new DMARTs?

And there's been a lot of discussion about the leflunomide versus methotrexate trials. This is the US301 study. This is the MN302 study. This is the 12 month data where numerically and at least statistically in MN302 there were differences between the two therapies.

These studies were, however, powered to show equivalency between active treatments, and when you look at the data over the two years in the year two cohort what you see, in fact, are very consistent responses and, most importantly, the differences between methotrexate and leflunomide in this study at one year and this study at two years are lessened, and so they become more obviously

comparable, and one could argue that two therapies which are equivalent will perform differently, one better in one study, one better in the other.

And the same may be shown also for the ACR 50s, and the point here is that virtually every treatment group, the ACR 50 responses, which are probably what we most want to see in our patients symptomatically, represent more than 50 percent of the ACR 20s in all of the treatment groups, and if we look at the ACR 70s, although they are really to small yet with our therapies to give us statistical comparisons. You can see that there's not a small number of patients who have really very striking clinical responses.

These are the responses over time and the HAQ disability index, again, in the year two cohorts between the three studies, the point being that patients begin with baseline HAQ disability indices of between 1.2 and 1.6, and they end up with HAQ disability index indices mean scores of 1.6 to 1.0, and whether that's MCID or more, it's clinically meaningful for sure, and I think we can

agree to that.

Finally, it looks very comparable to the data in the ERA study with etanercept and methotrexate in patients with early disease, 11 months of disease who would be expected to improve quite rapidly from baseline scores of 1.6 and HAQ disability index, and, in fact, they do, and this is maintained over 24 months, but we did not have the data to show the slide.

This is the ATTRACT study that we talked about earlier today, and again, this is an ITT LOCF study, but the point being patients remain on methotrexate in both of these treatment groups, but those who are receiving methotrexate plus placebo begin to deteriorate long term compared to the infliximab group.

We talked about health related quality of life and improvement in those domains which are different than just physical function or role physical. I think we can say that it's clinically meaningful if a group of patients now reach what are meant to be age and gender match norms for that

population.

And we see that also with the PCS score, with leflunomide, methotrexate, and Dr. Ware is performing a meta analysis of PCS, MCS and SF-36 data with arthritis therapies and have told us that this is the largest effect size he's seen in the PCS.

This is comparable data, again, with methotrexate and etanercept in the early RA study at 12 months.

And finally, although this is presented differently, this is data, again, showing clinically meaningful improvements in the PCS scores, in the infliximab/ATTRACT trial with active therapy on top of failed methotrexate.

So the results with leflunomide in terms of benefit, they're clinically meaningful whether MCID is the appropriate definition or not. The vast majority of patients are improved, and I think you would agree with me that these are comparable to improvements that have been observed with both methotrexate in recent clinical trials and also with

the biologic DMARDs.

Now, what can we say about risk evaluation? You've heard extensively about it this afternoon. So I will try to briefly highlight it, especially since no one is getting a break.

Quickly, the type of monitoring we do for methotrexate and leflunomide are LFTs, but also CBCs, and just to look at across the randomized controlled trials, Phase 3, you can see the percentage of AEs for CBCs and LFTs, SAEs in blue, and treatment related SAEs, and this is a profile that is at least positive for leflunomide compared with methotrexate and sulfasalazine.

At year two one might expect better tolerability. One sees better tolerability, but one still sees a positive profile for leflunomide compared to methotrexate and sulfasalazine.

What about rare adverse events? We talked a lot yesterday about lymphomas and so on. This is the incidence of lymphoproliferative disorders or lymphomas in the Phase 3 clinical trials with leflunomide, placebo, sulfasalazine,

methotrexate. This is per hundred patient-years, which represents .012 per thousand patient-years, and .020 per thousand patient-years for methotrexate and leflunomide. They are certainly not different, and this might be what we could consider the background incidence on our standard DMARD therapies in a disease that is prone to have development of lymphoproliferative disorders.

As you can see, also interstitial pneumonitis is represented only in methotrexate, reversible renal failure, again, only in methotrexate, and agranular cytosis only in sulfasalazine.

In terms of the safety profile then, I think you can agree that the year two safety profile is comparable in data that was presented both in the briefing document and discussed earlier, and basically we really believe by the controlled clinical trials that the serious hepatic adverse events are very comparable to methotrexate and sulfasalazine with the exception of one severe hepatocellular injury hospitalization which reversed

completely.

Withdrawals due to adverse events with leflunomide in these pooled trials really were quite comparable with sulfasalazine and methotrexate. There were fewer serious adverse events with methotrexate which were treatment related. There were fewer hepatic events than methotrexate. There were comparable serious adverse events with sulfasalazine and comparable hepatic events, and sulfasalazine in general is not thought to be as, quote, hepatotoxic, unquote, as methotrexate.

What did we learn from the post marketing surveillance? Well, there was a fair amount of discussion about the post marketing surveillance yesterday, but first I want to just say what is the world of, the universe of use of leflunomide.

Well, this is a rheumatologist prescribing and this is actually a physician panel data for prescribing use through December of 2002, indicating that there are approximately 294,000 scripts through 2002 in the United States, and of

those prescriptions written, 84.4 percent of them are written by rheumatologists. It's comparable we see for etanercept and anakinra. We can explain the differences with both infliximab and methotrexate, in part, because of the difficulty in tracking methotrexate use and also because of concomitant use of methotrexate with infliximab and its use in Crohn's Disease as a monotherapy.

What have we talked about about mean exposure time to leflunomide? Well, it is not four months or 4.5 months. In the Aetna study it was a mean of 19 months. In Fred Wolfe's database, it's a mean of 15 months, and in the Eisen data which has been published as abstract form and it's in publication now, it's 17.6 months.

And worldwide until approximately March of 2003, we could say that approximately 600,000 rheumatoid arthritis patients have received this therapy.

So if we look at reporting rates in terms of post marketing surveillance, it's very hard to define a rate, but one can take the Medra

(phonetic) terms as reported to the FDA, and one can take IMS data for prescription use and come up with an estimated denominator and try to come up with an estimated reporting rate.

It's agreed that this is only an estimate. It's not accurate, but it often can give us at least some comparisons that may be useful.

We're used to looking for hepatic failure, interstitial lung disease, serious skin reactions in part due to nonsteroidal use, as we know, with Stevens Johnson and TENS Syndrome. Vasculitis and lymphomas, of course, are thought to be part and parcel of both the disease and potentially its therapy.

So I will run through these very quickly simply to show you some patterns and not to try to say that we can generate significant numbers from them.

This is already what Dr. Holden has shown you for reporting rates for hepatic failure, and we know that although this rate is flat, we've seen that there are cases reported, in fact.

This is for interstitial lung disease. This is what we've seen with cutaneous reactions, and this is just through the end of 2001.

This is vasculitis.

And this is lymphoma as we've been discussing, and again, this is through fourth quarter 2001.

We've also realized that recently, even since the cyclosporin clinical trials, that we need to recognize hypertension as a comorbidity in our patient population; pancytopenia because of the associated marrow abnormalities from an infectious autoimmune disease that is chronic; sepsis and tuberculosis that was discussed yesterday; and demyelinating disorders which have become increasingly recognized.

And there are some interesting patterns here. This is hypertension. This is pancytopenia. This is sepsis and tuberculosis, and this is demyelinating disorders.

So in terms of a rheumatologist point of view, I think we could argue that leflunomide is

comparable to methotrexate without the known interstitial lung disease or the known reversible renal failure, and reports of these types of adverse events and other rare ones and ones that are increasingly becoming recognized certainly there may be some differences between leflunomide and the other new DMARTs, but they represent signals of potential risk, but they say that they're comparable therapies.

Spontaneous reports of acute hepatic failure are really rare. They are confounding factors that are very common, as we've discussed. The exact incidence is really unknown, and I think we could argue again that reported rates are comparable to the other new DMARDs, at least based on our surveillance data and these cohort studies.

Briefly, from the Aetna cohort study, the nested case control study and Dr. Fred Wolfe's national data bank, rheumatic diseases, we've seen a very similar pattern. Basically the rates of hepatic events observed with leflunomide were comparable to the other DMARDs, be they biologics or

be they in combination.

In the nested case control study, there did not appear to be an increased risk for adverse events that were associated with liver or hematologic or pancreatic adverse events or serious opportunistic infections or septicemia.

And in the national data bank, rheumatic disease, by Dr. Wolfe, in fact, the events for liver hepatic adverse events, comorbidities, hospitalizations, and liver biopsies, which in fact, are easy things for patients to recall in surveys performed on a six months basis, there did not appear to be an increased risk for patients receiving leflunomide versus those receiving methotrexate.

And Dr. Wolfe is available to discuss this in more detail.

Now, the estimates of serious liver adverse events range between a low or a high of one in 3,000 to one in 5,000, following Dr. Goldkind's very detailed and exhaustive review.

What are the background rates? Well,

they range all over the map here, too. I think we can say that in the context of what is occurring, there is a signal, but it is a signal that indicates that these events are very rare, and there is some evidence to say that patients with rheumatoid arthritis may have a higher incidence of serious liver adverse events, i.e., those that can cause a hospitalization.

Yesterday it was asked when this was shown about the lymphoma evaluations in the national data bank what this group of patients meant because they were the ones who were not receiving the methotrexate, infliximab or etanercept, and Dr. Wolfe was very nice last night to perform a brief back-of-the-sheet computer analysis for us.

I'm to point out that this is all leflunomide patients. So they may be receiving leflunomide in combination with any of these above. So it's not quite a comparable analysis, which is why it's in a different color.

But the point is that there is certainly, as I showed you for the randomized

controlled trial database, no increased signal for lymphoproliferative of lymphomas if we look at what's observed versus the relative expected rate and come out with this standard SIRs ratio that we were looking at yesterday.

So in summary, the RCTs, the pooled analyses of the RCTs would really say that leflunomide is comparable to methotrexate and to sulfasalazine. The post marketing surveillance and the nested case control studies and the national data bank for rheumatic disease would basically say again leflunomide is comparable to the other DMARDs. It's comparable to the new biologic DMARDs as well.

If we talk about the positive side of this, and that is the number needed to treat, and we have talked about this briefly before, calculated as a reciprocal of an incremental benefit, we go back to the patient reported outcomes. We can look at the HAQ disability index. We can look at the PET Top 5 despite its criticism as making the HAQ too complicated. The patients didn't appear to have trouble completing those case report forms on a six

month basis.

And the SF-36 PCS, we see very comparable results, and of course, what's very interesting is if you look at this data, you find out that the physical functions that are queried in the HAQ are important to patients in very different ways, and approximately 40 different lists of the top five functions come out when we look at this.

And we can see that for leflunomide versus methotrexate, as well as methotrexate versus placebo, there are benefits that are offered by these therapies.

Another way to quickly look at methotrexate combination trials' step-up therapy, Dr. Hochberg's analysis, and he's in the audience, too, if you want to ask questions. Based on the ACR 20, 50 and 70, when a DMARD or a DMART is added to failed background methotrexate therapy, we can see, again, that the positive benefit, low NNT values are quite evident for etanercept, infliximab or leflunomide, as well as Anakinra until we get to the ACR 70s, which are difficult to compare

statistically at any rate.

So the conclusion in my mind would be that leflunomide does provide significant and sustained improvement in signs and symptoms and radiographic damage; improves physical function over two years in those patients who can tolerate this therapy and stay in the trials, and this is reflected in all domains of health related quality of life.

The safety profile is comparable across two years of treatment in controlled trial settings, and the benefit-risk profile really looks very comparable to our gold standard, methotrexate, and the newer biologic DMARTs.

What's important is that each of these therapies has their own unique benefit-risk profile. We are rheumatologists need to be cognizant of that benefit-risk profile, but we've learned hot to monitor our therapies, and we've demonstrated that we can do that with methotrexate.

It appears that that type of monitoring is what's required for leflunomide, but in fact, it

has had that labeling since it's approval in September of 1998.

So all of these new therapies, including leflunomide, represent important treatments for this chronic disabling disease in a population where we still have very limited therapeutic options.

Thank you.

And now I am asked to go ahead and say that this concludes the Aventis presentation so that Dr. Ruth Day can have her moment in the sun at this hot podium.

(Laughter.)

CHAIRPERSON ABRAMSON: Okay. Very good. Thank you, Dr. Strand.

And Dr. Day will be presenting discussion of labeling rare serious events. Dr. Day is from Duke University.

DR. DAY: Good afternoon, everyone. I have a variety of comments about labeling issues, and the key concept is cognitive accessibility.

Cognitive accessibility is the ease with which people can find, understand, remember, and use

drug information, and of course, do so in a safe and effective way, and by people I mean both the health care providers, physicians, pharmacists, et cetera, and patients and caregivers.

Many cognitive principles underlie people's ability to understand labeling information. Here are just some of them.

Information load; how much information is too much? Yesterday we were talking about potentially adding a warning, and someone said, "Oh, there's already too much in there already. Don't put anything else in." So how much is too much?

Another principle is chunking, and that's basically about putting together what goes together, information about the same topic together.

Coding has to do with once you have a chunk to give it a name, to give it a title or a subtitle, and that helps people code it into memory, pull it out later, and understand it more thoroughly.

There are other kinds of cognitive principles we won't be talking about today. One we

will look at quite a bit is location. If you're going to add something to the labeling, where might it go?

The readability of the labeling; the ease with which people can actually comprehend or understand the information; the extent to which the labeling enables you to focus your attention on some information and filter out other aspects versus looking at a variety of things at the same time.

So there are a whole variety of cognitive principles that have been well studied in laboratory situations for many decades.

So let's talk about load. How much is too much? Ordinarily when people think about this in the context of labeling, they focus on information load. How many pages can we expect people to read and understand? How many words?

Well, it turns out there is no answer to that because what is important is not the information load, but the cognitive load.

Cognitive load involves the mental effort that's needed to read and understand and

remember information so we can look at the number of mental steps and the complexity.

So if we were going to add a possible warning, and I am going to put one up here; I am not suggesting it should be a warning on any label that we've ever heard of, but if we were to add a possible warning like this one, "rare but life threatening liver toxicity has been reported including acute liver failure," now this is a potential warning that some people might entertain for the current product that we're looking at today.

So if we were to entertain adding this to the label, the next question would be where should we put it. What is the appropriate location?

Well, there are a variety of possible places. Obviously the black box warning or the warning section, and there are reasons for putting it one place or another, but we had asked what would that look like.

So here is the current page 1 of the Arava label, and that's currently what's in the black box, and so it would be added to that, or it

would go later on. It's approximately page 7, something like that, in the warning section.

Okay. So it could be added in either of those two locations, for example. But we might say, "Does it matter?"

So that's the question I'd like to address now. Does it matter if you're going to put something in where you put it?

Well, in order to answer that question, we took an empirical approach in my laboratory and did an experiment to find out. The basic procedure is shown here on the display. So over time participants study the label for a sufficient amount of time, and then we ask them to perform a variety of cognitive tasks.

The content of those tasks includes things such as what is the indication and focus specifically on warnings, and we're looking particularly at liver failure, which is that added sentence there.

And the tasks include things like free recall, being able just to tell what all the

warnings were or some of the warnings were on the label, and then recognition where you give them potential warnings and have them say yes or no, whether it was contained in the label.

So here is where we actually did imbed that sentence, either in the black box or in the warning section, and I've provided that extra sentence for you here in red just to alert you as to where it was. It was not shown in red to the participants.

So now we want to look at results for the free recall experiment. Again, the question asked to the participants is what are the warnings provided in this label, and we're going to plot percent correct as a function of where they happen to see it.

On a random basis, half of the participants saw that added warning in the black box up front and half of them saw it in the warning section later on, and you might want to predict in your own mind which would be better.

And now that you've done that, let's

look at the results, and it might surprise you. People who got that added sentence in the warning section did much better than did the people who got it in the black box warning up front.

As a matter of fact, it was a two and a half times better percent correct in this experiment. The same data are now shown on the next slide showing the full range from zero to 100 percent correct in order to point out, and for some reason I have lost the -- oh, my gosh, my gosh, my gosh. Don't look. Don't look.

(Laughter.)

DR. DAY: All right. The data are shown here with the full scale from zero to 100 percent so that you can see the overall performance level is low. It is, but it's still two and a half times better for the people who saw it in the warnings section.

Now we'll go to the next experiment, the recognition paradigm. In a recognition paradigm, we have basically a fill-in-the-blank item, and we'll say, "Is such-and-such a warning that's provided on

this label?"

And over a series of what we call trials, we insert different things in there. So is malignancy a warning on the label? Is stroke a warning on the label? Is liver failure a warning on the label?

Let's look and see what happened just for the liver failure item. And there are the results. Again, the people who got the information in the warning section did better than those who got it in the black box.

Let me add in now this blue line which shows you where chance is. You might have noticed overall performance was high, but chance is 50 percent because it's a two response alternative. On each trial just say yes or no. All right?

So the black box performance is modest. It's in the middle range, in the 70s, and when it was provided in the warning section, it was over 90 percent.

Okay. So we have two different research paradigms that have given basically the same

results. So now the question is: does it matter? And the answer is yes.

The warning section location did increase the ability to remember the warning and recognize the warning. Why? Well, it seems kind of obvious. It's in different locations.

That's not the only story. There are other things going on here. Let's go back to that concept of chunking that I told you about before. Put together what goes together and separate it from other things.

So let's go back and look at how we added the sentence into the black box and the warning as well, the warning section. You'll notice that the new sentence just picks up where the last sentence ended. Okay? And all of those sentences before it are about pregnancy, and then this is about liver toxicity. And there are precedents for this in labeling.

Okay. Another way to do it would be this way, to leave a space between all of the pregnancy warnings and then have a new space for the

liver toxicity.

An even better way would be to do that and then do not only chunking, but coding. Give a name to each one of those chunks of information.

So it isn't so much a black box is a black box is a black box. It's how you present it that's going to make it more or less effective.

Here are just some other examples of other drugs currently available and what their warning sections look like. This one goes on and on, puts everything together. This one chunks things into hepatotoxicity, pancreatitis, et cetera.

And so I would like to just tell you that there are a huge number of experiments that show that when you chunk information and give it to people, they do much better with it. They can find it, understand it, remember it, use it to solve problems in the future much better.

There's over a half century of research that says that chunked information is better processed than unchunked. Similarly for coded versus uncoded information, give it a name so people

can understand it, store it away, and then retrieve it later when they need it.

So now, there are other issues. There's legibility, and I'm not going to talk about font size, but I would like to address the notion or the fact of capitalization. There are studies that show that if you capitalize information, it's good for warnings, but it's good for warnings only when it's a word or a phrase, such as "stop" or "no admission." All right?

It is not good for text. People cannot read text when it's all capitalized, and I do research in my lab not only on drugs, but medical devices and with real patients, with college students, with professionals. People complain they can't read the capitalization.

So here is the same black box warning now in the upper/lower case which facilitates reading text, and there are examples of this in the PDR for approved drugs as well, and there's one example.

So now another issue is readability.

Well, going back to the current black box warning for Arava, you'd say, "Well, what's the problem? It's only 48 words and three sentences. Our physicians are smart people. Patients who are motivated enough to take a look at this thing aren't going to understand this."

Well, it turns out that 66 percent of the verbs in this little passage are passes. There's a huge amount of literature in Cycle Linguistics which shows it's harder to process sentences in the passive voice. It takes longer and you're more likely to make a mistake in understanding it.

The grade level is 12, but that's really an under estimation because there's a cutoff in that score. It doesn't go any higher than 12.

And furthermore, there is a problem about readability. Readability is not the same thing as comprehensibility. Readability is not the same thing as comprehensibility.

What is readability? There are lots and lots of different measures out there. They all use

two types of things, that is to say, word familiarity. How frequent in the language are the words in the sentence and sentence length, number of words per sentence? That's all it is.

So there are ways to artificially bring that readability level down to some nice level, and especially in patient materials, say in med. guides or other kinds of things that are oriented specifically to patients.

You can manipulate and bring the readability level down to whatever your target is, sixth grade, eighth grade, whatever it is. That does not insure comprehensibility.

For comprehensibility we have to look at the number of propositions or idea units packed into each sentence because that can overload cognitive processing, and then also the syntactic or grammatical complexity and other factors as well.

So let's look just a little bit at linguistic structure. Here is the current black box warning for this product, and I've put in red all of those extra little words, mostly prepositions.

So the first sentence is, "Pregnancy must be excluded before the start of treatment with Arava." Not too bad, but let's go to the last sentence.

"Pregnancy must be avoided during Arava treatment or prior to the completion of the drug elimination procedure after Arava treatment." That is hard to process, which brings up the whole notion of lard.

(Laughter.)

DR. DAY: Lard is extra words in a sentence that make it difficult or hard to extract its basic meaning. There is a gist or a basic meaning in a sentence, and extra words can make it difficult to get at it.

And there is a de-larding procedure. You can rewrite --

(Laughter.)

DR. DAY: You can rewrite each sentence using only essential prepositions. Prepositions do exist for a purpose in the language, you know, but use only those that are essential, and full verbs.

So get rid of the "is" verbs wherever possible, such as passives and the situations, and make the verbs have action in them.

So here is an example. "This sentence is in need of an action verb," is a "lardy" sentence, and if I de-larded it, it would say, "This sentence needs an action verb." So those of you who do have the handout, there was a slight typo. The N is not there. So "this sentence is in need of an action verb" goes to "this sentence needs an action verb."

Okay. So let's go back to sentence number one in the original. It would go from "pregnancy must be excluded before the start of treatment with Arava" to "exclude pregnancy before starting Arava treatment."

Okay. So I de-larded the whole thing, and there are different ways to do it, but the original versus the de-larded version, we can now compute the lard factor.

(Laughter.)

DR. DAY: The lard factor is simply the

number of words in the original minus the number of words in the revision divided by the number of words in the original, and we saw formulas like this yesterday for other purposes.

(Laughter.)

DR. DAY: When you do that, the lard factor for the current Arava box warning is .23. That means there's about one quarter of the words are extra words that are going to make it more difficult to pull out the meaning.

Now, why should we care? If you really work, you can understand that box, but there's so much in there. You have 40 patients sitting out there. You've got to work with this one, et cetera, et cetera. So there's a problem of mental economy here.

And if it's so difficult to dig out what you need from labeling, people are going to go to it less and less and problems can happen.

Okay. So there are many other experiments I could talk about on readability and attention and comprehension, memory, problem

solving, decision making.

In the interest of time I'm not going to throw anymore research reports at you and anymore numbers, but getting back to our results here today, the overview slide that I showed you before, we can now answer the why question a little bitter.

Why do we get those results? Because we made a box warning. We made the information in a certain way so that location was certainly relevant, but also chunking, legibility, readability, comprehensibility. And if we can just enhance all of those, we could probably put it lots of different places, and it would be attended to, remembered, and understood more readily.

So a black box can, indeed, be effective, and I lost my number one there. I don't know why. And a black box will be effective when it's legible and it's not all capital letters, when it is chunked by type of warning, and when those chunks are coded, titles for those chunks.

And of course, there are advantages to both the black box and the warning section for this

type of information. The black box is great. It's up front. There's a tremendous amount of information that shows.

If you give people a whole long set of information, they're most likely to get the beginning and the end, but they lose stuff in the middle. This is called the serial position effect. So it 's up front right where you have people, and they're going to do well with that.

It's also in a box. It's visually distinctive, and furthermore, it serves an alerting function in this kind of document which we all know about.

There are advantages to putting things in the warning section. There's the context of having all the warnings together and also the specific types. There is a whole section of hepatotoxicity.

So let's step back from all of this right now and talk about information in labeling. Labeling serves a lot of purposes. It serves a regulatory purpose. It serves a legal purpose, et

cetera, et cetera.

And when people are developing labeling there are a lot of reasons to put in a lot of things,a nd often the tendency is, "Oh, let's put that in and let's put that in. Let's cover ourselves," and so on.

So let's say we had idea labeling where every possible thing that could happen is in there and everything else is good and correct to the best of our knowledge. So everything would be physically present.

However, it could all be physically present, except it could be functionally absent. That is to say if it is not presented with sufficient cognitive accessibility, people are not going to be able to notice it, find it, understand it, remember it or use it. So it is functionally absent.

So I'm arguing here for evidence based labeling. Probably when I said "evidence based labeling" you thought of, "Oh, yes, let's put in all the data from clinical trials, post market

surveillance, et cetera."

I'm also suggesting that when there are questions about the effectiveness of certain language and location and so on for labeling, that label comprehension is a good thing to do.

We can get empirical evidence for the effectiveness of adding warnings and so forth. Now label comprehension is involved in over-the-counter drugs these days. So that's a regular part of studies, and it is not required for prescription drugs, but when questions like this come up, we really can get some evidence.

So what usually goes on? Well, we look at everything that has to go in, and we have this target in mind of what's got to go in the labeling. So in the little cartoon here, there is the target with the folder of all the stuff that everybody has ever collected that might to in the labeling, and then it comes down to, well, what can we put in? And should this go in? And how should we say it? And is it too much?

And then in the end, although many

wonderful decisions are made, sometimes we can actually be a little bit blindfolded and just say, "Well, let's put it in just in case."

That doesn't need to be the case. We can, indeed, get empirical evidence about these labeling issues, and so if we get empirical evidence, we can then enhance our labels, and they can be more effective.

Thank you very much.

CHAIRPERSON ABRAMSON: Thank you very much, Dr. Day.

Two of our members have to catch an airplane to the middle part of the country and both of them are known for having no lard in their answers.

(Laughter.)

CHAIRPERSON ABRAMSON: So I would ask -- we're going to go to Question No. 1 on the safety issue, and having heard what we have on the leflunomide benefits and hepatotoxicity, I'd ask first Dr. Williams and then Dr. Brandt to address the first question.

Considering the universe of available disease modifying therapies, is the benefit to risk profile for leflunomide acceptable for its current indications?

Dr. Williams?

DR. WILLIAMS: My answer would be yes. I consider it analogous to methotrexate in both efficacy and toxicity, and I think it's a valuable addition to the armamentarium.

CHAIRPERSON ABRAMSON: Dr. Brandt.

DR. BRANDT: Yes. Same reasons.

PARTICIPANT: No argument here.

CHAIRPERSON ABRAMSON: Good.

(Laughter.)

CHAIRPERSON ABRAMSON: And I may just put this question to the committee because the meat of our discussion, I think, is more on Question No. 2.

Does anybody disagree among the members of the committee with the answers of Dr. Williams and Dr. Brandt?

(No response.)

CHAIRPERSON ABRAMSON: So we have a consensus that I think is clear in terms of the risk-benefit of this drug, that the data, all things considered, appear comparable to the other DMARDs that patients are offered, and all drugs have their different profiles, but there's an acceptable benefit-to-risk profile for leflunomide.

For the FDA perspective, is that --

DR. WOODCOCK: Yeah. I would ask that you ask the hepatologists to comment on the totality of the data, on the liver toxicity.

CHAIRPERSON ABRAMSON: Yes. Definitely I was going to focus more on that in Question 2.

DR. WOODCOCK: Fair enough.

CHAIRPERSON ABRAMSON: And if there's a serious difference from that that emerges, we can refocus on that. But for the moment, going from Question 1 to 2, I think that's the consensus on the committee, and then let's look at the liver toxicity in the sequence of the questions if that's okay.

All right. So if the answer is number one, what labeling or other communication of risk or

risk management is warranted for optimal safe use of leflunomide?

And I think this is going to take a lot of discussion and involve the hepatologists. I would just ask because of the plane situation that you have a comment or two from Dr. Brandt and Dr. Williams, if they want to say anything because they may not be part of the more extensive discussion.

DR. WILLIAMS: I would like to just say that with most of these disease modifying treatments that we're using to dealing with toxicity as rheumatologists that I have not seen anything presented here that was surprising that was not already being monitored for. I would not think that any labeling change would be necessary unless it was to improve readability as Dr. Day has suggested.

CHAIRPERSON ABRAMSON: Okay, and, Dr. Brandt, as a preliminary comment?

DR. BRANDT: I think with respect to content, I think we're okay the way we are.

CHAIRPERSON ABRAMSON: So I think now we should go back into this issue of the liver toxicity

and get deeply into that.

DR. KWEDER: Excuse me. Dr. Firestone, we actually would appreciate it if on Question 1 if you're done with Question 1, if you could take a formal vote of the committee. That would be very helpful, a yes/no.

CHAIRPERSON ABRAMSON: Okay. So why don't we go around the table?

Yes, Dr. Day?

DR. DAY: I think some of us would be better able to do that vote once we've heard from our colleagues on this issue, our liver specialists.

DR. KWEDER: That would be fine. We just want to make sure that we get a clear answer.

Thank you.

CHAIRPERSON ABRAMSON: We do have both Dr. Seeff and Dr. Lewis with us and would like to hear what their thoughts are about both the adverse event reports and the other information that we've heard today.

DR. SEEFF: I'm going to try to keep lard out of it, but I may not be able to.

I came here with a slightly different view, but I am compelled with the data that I heard today.

On the other hand, I think there's a broader issue than just what's happening here with this particular drug, and that is I don't believe that we really know how to make a specific diagnosis of drug hepatotoxicity. We are dependent upon surrogate markers, the surrogate markers being enzymes, amino transferase for hepatocellular disease, alk.phos. (phonetic) for cholestatic liver disease, perhaps suggesting that this may be a hypersensitivity reaction, the so-called Hy's view that jaundice is what's the cause of this.

And, by the way, let me just tell you that I have been a hepatologist for almost 40 years, and I started working with the eminent Hy Zimmerman in 1964, and I was with him all the way through until he died. In fact, I was with him when he died.

I'm very angry with him because if he were not dead, he would have been here instead of

me, and I wouldn't have to go through this interrogation. So --

(Laughter.)

DR. SEEFF: But I am concerned that we don't know how to make a diagnosis, and I say that because the reason why I have changed my mind is that the data that I was given were not the same as what I heard today. What I got were the Medwatch forms, and the Medwatch forms as I understand them, at least what I looked at, are absolutely or not absolutely, but largely meaningless. There's just not enough data in there to be able to make a definitive diagnosis one way or another.

The data are not in there. There's a lot of information that is missing. One of the things that I've actually mentioned to Dr. Goldkind that I think is seriously missing and that we really have to begin to think about for the future is the fact that there are other products that people are now taking such as the alternative medicines and herbal products that may, in fact, be responsible for some of the hepatotoxicity.

In fact, I have seen a number of cases now in which this has occurred, but unless one actually asks that question, you don't know about it because people are reluctant to talk about it.

So I think we have missing information that would be very helpful in trying to define this. I came away with -- I was sent four groups of Medwatch forms. They were called acute liver failure. I can't remember. Severe liver disease. Some were from the United States; some were from Australia, and the question that I was asked was is this definitely not; is this definitely yes; is this probable or is this possible, and I had to come away with what I had to say that some of these cases were possible based on the information that I was given and the ability to try to understand what's going on.

Now it's easy enough to say, well, you know, the patients were on other drugs that may, in fact, have been implicated. But, on the other hand, if you were somebody who owned the other drugs, you'd say, "Well, it was the leflunomide that was

implicated and not the other drug."

So, yes, indeed, it could be. We don't know which it is, if indeed it is associated at all, and so I think that this becomes a real problem, particularly when you have multiple drugs because there is no definitive way that I'm aware of, although I'm in the presence of some outstanding hepatologists and people who are much more expert than I am in hepatotoxicity, who may, in fact, give me the information, but I don't know specifically how to diagnose hepatotoxicity other than basing it on temporal relationship between the use of the drug and the development of abnormal enzymes, and that may or may not be correct.

The second thing that I think we need to think about, and I think that this also transcends the discussion here, is the fact that we do know that there is elevation of liver enzymes not uncommonly, but there appears to be a distinction between elevated liver enzymes and hepatotoxicity because sometimes the enzymes go up, stay up at a modest level, and may stay like that for a long time

or go down despite the fact that you go on using the drug.

We assume that that is absolutely benign, and it may well be, but let me remind you that there are two parts to liver disease that we are concerned about. One is the acute problem: fulminant hepatitis, patients coming into the hospital because they jaundiced, and so on and so forth.

But there's a second part to liver disease, and it's a most important part of liver disease, and that's the potential of chronic liver disease, fibrosis. I think that actually in my view the most important thing that we have to study and research in liver disease is how to define fibrosis without having to do liver biopsies because almost all liver disease which is chronic, chronic liver disease, is something that may be associated with progression to fibrosis.

I mean an example is so-called non-alcoholic steato hepatitis, NASH, that we sort of set aside for so many years, is meaningless. Well,

NASH is no longer meaningless. We've got a big study at the NIH, thousands and millions of dollars being spent on trying to understand NASH, and why? Because we think that these people may be the people responsible ultimately for so-called cryptogenic cirrhosis and potentially even hepatocellular carcinoma.

Hepatitis C, the big problem is not acute Hepatitis C. It's chronic Hepatitis C, and it's not chronic Hepatitis C per se. It's advancing fibrosis. People die only if they have cirrhosis largely. Well, they die from obesity. They die from diabetes. They die from too much drinking.

But if it's liver disease and they have Hepatitis C, they're going to die if they cirrhosis either from end stage liver disease or from hepatocellular carcinoma. So I think evolution to chronic liver disease important.

Now, I am not suggesting that this has anything to do with what we're doing here, but I think that we should begin to think about the possibility that if we're using a drug that is going

to be used chronically and may lead to chronic elevation of serum enzymes, that we should not necessarily discard that as meaningless. I think we need to consider the possibility of studying such things before we say it doesn't have any meaning.

The other thing, of course, is that when you have multiple drugs, which is the case over here, what we looked at, these are patients on methotrexate and on Celebrex and on leflunomide and so on and so forth. Which one is it?

And there's no marker which says that it is A or B or C or D. So it's a real problem, and I think that one of the things that the FDA is constantly faced with and ultimately we're going to have to do something about is to learn about better markers of hepatotoxicity, you know, whether the micro arrays and identification of genes that may be responsible for defining serious liver disease, and the ability to identify those genes becomes one of the ways of doing this or not I don't know, but this is lard. I understand.

But all I can tell you is that I did

come away --

(Laughter.)

DR. SEEFF: -- with a few that based on what I saw there were some cases that could conceivably have been a consequence of leflunomide.

On the other hand, what I heard from Dr. Goldkind as part of the FDA presentation and from what the Aventis people had to say, it really has not been associated with severe liver disease, and I think that that's compelling data.

I personally would have liked to have had more information on all of these patients. I would have liked to have had the charts. I know that you don't have it.

I also know that the problem is that people don't gather that information. I tried when I wrote my letter to you to say what would be needed if we wanted to identify hepatotoxicity, and there's a series of events that every one of us know about. We would need baseline enzymes. We would need to follow them with enzymes. We'd need to stop the drug and see what happens if the enzymes go down,

and so on and so forth.

There's a whole series of things, and that was not available.

I would have been more comfortable though had I had more data, had I had the actual charts, and had I had a chance to look at that to say that these were definitely not or that these were definitely something else.

So I concur that there is no evidence on the basis of what I learned today that this drug is associated to any great degree with acute liver disease.

I remain uncertain about whether there is chronic enzyme elevations that are worth looking at and perhaps following up on. I don't know whether these people have had subsequent liver biopsies, for example, to see whether they develop fibrosis. We know that Hepatitis C takes 20 years before you end up with fibrosis or cirrhosis, and I don't know how long leflunomide is going to be used.

I am compelled that this is very good. I was extremely impressed with Ms. Leong's

presentation because I think that one of the things that we do have to take into account in my view is the severity of the disease.

If the disease is so disabling, as we heard from her, it's worthwhile using a drug even if there is hepatotoxicity, and I think then the physician is more likely to use it and the patient is more likely to accept it.

In this case clearly people with severe RA deserve to be treated with the best possible treatment, and this is at least as good as and perhaps, with not being a rheumatologist may be a wee bit better. The hepatotoxicity, as I say, seems to be not a major issue.

But I think that the FDA with all due respects needs to sit down maybe with the NIH, maybe with other people, and try to think through more about how we assess the issue of hepatotoxicity and what better ways we can devise in order to identify hepatotoxicity distinct from viral hepatitis, from autoimmune hepatitis, from alcoholic hepatitis.

Even though there are many clues,

sometimes it's very difficult and I'm very concerned. I'm particularly concerned, for example, in people with cancer with multiple drugs.

I know that I'm off the topic, but I'll stop at that point.

(Laughter.)

CHAIRPERSON ABRAMSON: Thank you very, very much.

Dr. Lewis, do you want to comment as well?

DR. LEWIS: Well, as another graduate of Hy Zimmerman University. I mean I share many of the same thoughts that Dr. Seeff elucidated.

We need to address the issue for the committee though. Was a signal identified in these spontaneous reports?

And I think it was in a sense that if you're got, you know, 80 reports or however many it was, that that's a signal.

Now, what's it a signal of? It's not conclusive, but it means that you got about the business of looking into these cases, which has been

done, and you come up with an assessment of what do these cases all mean.

And our reports are here in the briefing books, and I, too, would have liked to have had all of the information on these cases, and in fact, the ones from Australia, I think, virtually none of them had any significant data provided.

We've sort of been hacked to pieces here this morning, you know, with no pun intended. Why can't we get decent data about real important safety issues? And it would be a complete remodeling of the spontaneous reporting system, I know, and lots of people are concerned about reporting for lots of reasons. There's medical legal concerns. Maybe we have to indemnify anybody who writes a Medwatch report.

But I'm also struck by the fact that just because somebody sends in a report and it's a very serious alleged reaction, if they can't take the time or provide us with full information on that kind of death or liver failure or renal failure, whatever it's going to be, how important was that

report and how convinced was that reporter that it truly was the drug and nothing else?

And we have a conundrum a little bit because I sit here as a clinician, and if somebody is on multiple drugs and has enzyme elevations, which I see every day, I have to make a judgment about what did it.

And I can sometimes delve back into the record. I can ask for more information. We can't do that here for many of these cases, although I certainly know it's possible to go to the reporting physician or whoever it was and ask for more information.

Because impugning a drug with circumstantial evidence means that the patient is not going to benefit from it any more. They're going to be off of it. We often may not continue to look for what the real cause of the injury was, and I think it confuses our safety profiles. We now say we've got all of these cases of liver failure and everybody just takes them at face value, which you can't do.

And what we attempted to do in our analysis was to the best that we could with the information is give you our opinion, and a very few of them I concluded were possibly related. I didn't think any of them were definitely related based on what I could tell.

It begs the issue though of the ones who are so inadequate as to what do you do with a very serious allegation and you've got no information at all. And in my experience, which I've already touched on, if you've got no information to back it up, if there's nothing in the literature and there's very little, if anything in the literature on any spontaneous case reports of liver failure with this drug or anything else to look at after several years of being on the market.

I have to wonder whether or not that absence of real information is just that. It's because it wasn't related in some way, and that's sort of how I have to interpret it.

So for the committee's point of view, I agree with Dr. Seeff that I'm swayed by the evidence

with all the data mining techniques that were used that to me there's not a signal that jumps out and says that this is going to be another troglitazone.

I think we would have seen that already, you know, with the length of time it has been on the market, and in fact, the two of us have reminded each other that four years ago almost to the day we were here discussing whether troglitazone remains on the market for another year, which it did with no further deaths with the appropriate monitoring and whatnot.

I guess the only question for the committee, and it's really going to be from the FDA's point of view: does the labeling stand as is?

We've already heard, you know, acute liver failure or possibly fatal liver failure. Should that be added to the label? If any one of these cases is so convincing that we think it's related, the death might be related to liver failure, I think an N of one could be in the table. I mean, is that a -- but, again, it goes to the risk-benefit, and I think that the benefits outweigh

the risks certainly in terms of liver toxicity.

CHAIRPERSON ABRAMSON: Thank you.

Before we get to a discussion of the label, I would like to get some other people's opinion on the adverse events. I know, Dr. Makuch, you had written a letter as well. I'd like to get some initial feedback from people before we reach a consensus about the label.

DR. MAKUCH: I don't have much to add. I certainly respect and agree with the two individuals who just spoke.

I think, you know, my comments are probably oriented a little bit differently, and that is I think that the Office of Drug Safety undertook what was a signal, and I think they undertook an effort to investigate that, and they came up with a modeling procedure.

In the letter that I wrote, based on my review of that document, one of the things I suggested, and I was unaware of all the studies done until sitting here today, that their modeling procedure then be validated against actual data, and

I think that today's information presented here gives a very useful tool, in fact, for validating or not validating the model.

But, again, with the information that they had at that point in time, I think they undertook a good effort, but in the end I believe that the data in all of these studies I think give a very consistent picture of not a great concern with respect to this issue.

CHAIRPERSON ABRAMSON: And I would just comment having also written a letter, to pick up a comment of Dr. Seeff, is that we were asked to say if something was possible or probable, probable being there was no other concomitant medicine that might be implicated, and that was in a time frame that could be leflunomide.

So these decisions were very arbitrary, and in point of fact, given the absence of robust information it has the potential to overstate what a person really believes is a causal association.

So I think it's important even in a public hearing to make sure that people understand

when we might write possible or probably in response to the Office of Drug Safety, what, in fact, the conundrum that the reviewers are put in applying some of the criteria.

PARTICIPANT: Well, I actually just want to comment on that.

DR. GOLDKIND: I just wanted to say that we wanted in forwarding those cases obviously to leave you unbiased, not to try and lead you in minimizing or maximizing and to welcome you all to the world of post marketing surveillance.

(Laughter.)

DR. SIMON: But we also wanted to insure and open hearing of all the opinions. So we tried to give you exactly what each consult provided, including the ODS concepts so that everybody had the chance here to review all the potential opinions regarding what this evidence might mean.

So that's one of the reasons why we burdened you al with such extensive reading opportunities.

CHAIRPERSON ABRAMSON: As long as the

caveats are noted.

Dr. Gibofsky and then Dr. Fries.

DR. GIBOFSKY: Dr. Simon actually opens the door to a concern, addressing a concern that I have, and that's a concern that's been nagging at me since Dr. Wolfe's comments, and that is that his opening comments almost cast a pall on the agency and on these proceedings.

The comments about the agency will be addressed by Dr. McLelland if he so chooses, but the suggestion that the proceedings here are somehow tainted by the absence of individuals who wrote a report and the absence of our opportunity to have a colloquy with officers of the agency who may have differing viewpoints is a concern because it suggests that my participation is somehow as an unwilling aider and abetter of a sweatshop, as it has been alluded to.

And that's something that I take very seriously. So, Mr. Chairman, I would like you to invite if any agency officer is here with a conflicting viewpoint to Dr. Goldkind's or sees the

evidence a bit differently and would like to present that before we reach our conclusions. I'd be interested in hearing that because I think it's appropriate for people to look at data sets differently and come to different conclusions, and the appropriateness of our decision has to be based on the synthesis of those different points of view.

CHAIRPERSON ABRAMSON: I would agree. The notion of a fair hearing is one of the objectives. If there is somebody who would like to comment, address Dr. Gibofsky's comment, I think we would be open to that.

(No response.)

CHAIRPERSON ABRAMSON: Okay. If not.

DR. FRIES: Thank you.

I want to drift slightly, but I think in a relevant way here. We're obviously very close to a group consensus, and we'll formalize that in a little bit, but I wanted to go back to Ruth's comments because it seemed to me that they hit in a way very relevant to the decision that we have here and also to a broader problem that probably Mark and

other people at the FDA should be considering.

And I call it in one sense -- there are several aspects of it that come home to me, but one of it is the problem of the false positive, and this is very important to us to recognize, that a false positive signal does harm. It keeps people out of studies. It keeps people off of drugs, our patients, that would be good for them because they don't like a particular thing that they've read or that they've read in the past.

With a colleague some years ago I wrote a science editorial called "Informed Consent May Be Hazardous to your Health," which I pointed out this and some other areas about unreadability, fearfulness, false positive types of things and implied some things that weren't in that piece.

For example, it has always bothered me that the PDR has so many things that didn't differ from placebo. Now, that's a way of, I guess, larding up the description because, in fact, you did studies and you showed that there was no difference from placebo. So there's no signal. So why worry

about this?

Or at the very least you would want to chunk these into alleged but unproven or something that was at some different level of certainty so that people could actually read in an informed, well written, de-gassed or de-larded way what the problems with this drug are, and they could understand it and recall it, and we could do it in between patients on our desk or we could pick it up on the palm pilot and actually get through with it because there are several principles -- and I'll just mentioned one that Ruth didn't mention. We were just chatting about it, but there also are some other rules.

She was keeping her transformations within the data that's in the current label exactly, but if you actually look at that, you find that some of the ways in which we write for patient comprehension just aren't there because you're not supposed to ever tell somebody to do something and then not tell them how to do it.

So the last sentence of this one was,

"Avoid pregnancy during Arava treatment and after treatment until completing the drug elimination procedure."

Well, I would say that's inadequately de-larded. What you want to do -- and it's inadequate. You want to say, "Avoid pregnancy during Arava treatment and for eight weeks thereafter." You have to give them some -- "drug elimination procedure"? What does that exactly mean? How do you incorporate half-life into that?

I mean how is a physician going to understand something like that? You have to give them the thing you want. You're going to base it on data, and you say for eight weeks or 12 weeks, whatever you decide to say, but say what the drug elimination procedure is so that people can understand this.

And I really think the people here should go back to Mark McLelland and consider the question of whether a very, very useful thing to do would be to have a half as thick PDR which didn't have false positive signals in it and was readable

by everybody, lay people included, and systematically have an approach applying some of these principles to the communications that go out as our warnings.

DR. WOODCOCK: Yeah, can I just comment very briefly on that?

Yeah, we are doing that, believe it or not, and we hope we have to do it through regulation, which is a stately process, but within the year you should see a new label that uses modern, to some extent, cognitive principles and provides us with an opportunity to move forward even better in the future.

CHAIRPERSON ABRAMSON: Dr. Day.

DR. DAY: I was going to comment on that part also. The proposed rule for physician labeling, if and when this comes out, is going to have a highlight section up front so that you get the latest information. It's going to focus people's attention, et cetera.

So there are a lot of things going on within the agency in order to achieve this.

I'd like to make a comment that Jim was just saying that that sentence was inadequately de-larded. It was de-larded, but chunks of information were left in because they were there from before.

So once you de-lard, you can see what the chunks are and decide whether they are adequately described and whether more information is needed or less.

My final comment is about the somewhat maligned Medwatch program, and I would like to say something positive about it. It is hard to get people to report, and you have to remember everything that's going on that make it difficult to report.

So say, for example, a physician has a patient who then has an adverse event of the type we're talking about. The form that is used is the same no matter what the adverse event is across any indications, et cetera. It is one form.

Correct me if I'm wrong. So it's one form.

So it cannot ask for everything that

would be needed for hepatotoxicity with all of the enzymes, et cetera, et cetera, and then for some other indication and set of drugs and so on.

So what they've tried to do is have one form fits all, and of course, it's not going to totally fit all. So I would not interpret the lack of needed -- I appreciate the lack of needed data in order to make a determination as to whether there's a signal from these case reports.

However, I would not conclude that the absence of the needed data is because the people didn't care enough or they weren't convinced enough, et cetera.

Sometimes physicians read in the newspaper that a patient has expired, and then they may remember treating the patient, et cetera, and write up a little follow-up thing, and this may have been some days afterwards and they don't have any of the data from the hospital experience or whatever.

So Medwatch is not perfect, but it's certainly better than nothing.

CHAIRPERSON ABRAMSON: Thank you.

Mary, two other members have to make a 4:30 plane, and I'd just like to ask for any -- well, they have to leave to make a plane. That's larding up this discussion. They've got to leave at 4:30.

Dr. Manzi and Dr. Seeff, I'd just like to ask if you have any comments before you leave that you'd like to have recorded in the discussions.

Dr. Seeff.

DR. SEEFF: Yeah, I have to leave in about five minutes.

You know, I was just telling Jim that we were seeing cases, and these cases were listed as acute liver failure from both here and abroad, and some of these I was uncertain about, and while none of these appeared in the databases that we were given -- that was the thing that concerns me a little bit because have these all been looked at and, in fact, all of these excluded and all of these said to be absolutely not acute liver failure associated with leflunomide. It must have been. Otherwise they should have been in one of these

databases.

But somehow or other I have a feeling that I still would like to see more information if I can on some of these cases because some of them I said it's possible, and of course, the possible was because there were other drugs and, of course, other drugs that could have been implicated.

But it's just as likely and it's possible that the implication was this drug and not the others or perhaps even the combination.

So while the database that I heard was so compelling and all of this seems so wonderful that there is really nothing to show acute liver failure. These were sent to me. I mean, I didn't make them up. They were sent to me, and they were actually listed as either serious liver disease or liver failure, and going through them, I was unable to be absolutely certain that it was not.

Now, I know this is a story about proving the negative, but you know, the fact is that I continue to agree with what is being said with some niggling misgivings, and if I had an

opportunity to look at these cases in more detail -- I don't want to do it because I don't have the time to do it -- but I'd love to see this done. I mean, I would just like to learn more about some of these cases.

But otherwise I won't change my mind, and with that, I will thank you and have to depart.

Thank you.

CHAIRPERSON ABRAMSON: Dr. Manzi, do you have any comments?

DR. MANZI: I really have nothing more to add, except to just compliment the agency. I think the sponsor for very thorough homework that they did in following up issues of safety, and with all of the data presented, given the limitations of everything, I feel perfectly comfortable with the risk-benefit discussed.

CHAIRPERSON ABRAMSON: Dr. Raczkowski first and then Dr. --

DR. RACZKOWSKI: Yeah, I just wanted to say that the agency did make -- the Office of Drug Safety did make extensive efforts in terms of

follow-up for all of these cases. Our safety evaluators spent a lot of time trying to contact the original sources, and so the case reports that were received by the consultants represented pretty much all of the information that we were able to gather, despite extensive follow-ups, particularly by Dr. Banelle.

I wanted to thank Dr. Day for some of her comments about the Medwatch program and AERS. I do think that AERS is a very useful and important signal detection tool.

I am a little bit concerned about some of the discussion here because I do think that it's clear that all cases that are reported to AERS aren't necessarily associated with the drug, but conversely, just because there's confounding factors doesn't mean that it's not associated with the drug.

And I think that much of the disparity that we saw in terms of the case evaluations had to do with, you know, how these confounding factors were faced and how they were addressed.

And I actually wanted to briefly talk

about two of the cohort studies that were done, and this came up a little bit yesterday about the difficulties with some of these cohort studies.

On the one hand, when you see numbers such as 40,000 patients with rheumatoid arthritis are enrolled in a study or 90,000 patients, it's very impressive, but then you whittle it down and you see the actual number of patients who are actually exposed to leflunomide, and Dr. Goldkind showed a slide of less than 3,000 patients in both of those studies.

It limits your ability to detect adverse events. Moreover, that 3,000 number doesn't reflect how long the individual patients stayed on leflunomide.

So I don't know if we have the data here or if the sponsor has it, but I think it would be interesting to know whether of those patients in those studies, how many stayed on leflunomide for six months or a year or two years so that we could get a sense of the ability to rule out an adverse event, let's say, one in 1,000 at six months or one

in 1,000 at a year, that sort of thing.

So I guess that's question number one, and the second question I had is the sponsor also showed a slide saying that based on those studies that toxicity was similar between leflunomide and some of the other drugs, and I wonder if the sponsor would comment on the ability of those studies to detect differences given the small sample sizes of patients who were actually on leflunomide.

CHAIRPERSON ABRAMSON: Okay.

DR. HOLDEN: In response to your first question, in the Aetna study there were actually over 5,000 leflunomide exposed patients accounting for over 11,000 person-years of follow-up time, and in that study, we estimated that the mean exposure time or the mean time on drug was approximately a year and a half.

DR. RACZKOWSKI: Right. I know you showed the mean data, but do you actually have the distribution?

DR. HOLDEN: No, I don't have the distribution.

DR. RACZKOWSKI: All right. Because I think the distribution would be perhaps more telling than a mean exposure time.

DR. HOLDEN: The second part of your question is a power kind of question.

DR. RACZKOWSKI: Well, in one of your slides you had indicated that based on the results of these two cohort studies, that the adverse event profiles were similar, and I'm just -- I wonder if you would comment if you think that the studies were actually powered to be able to detect realistic differences between rare adverse events.

DR. HOLDEN: Well, we knew going in that these studies would not be powered -- any database currently in existence is not powered enough to look at differences in very rare hepatic events or any kind of rare event. So we did not do power calculations prior to doing the study.

And of course, after we analyzed the data, we'd look at confidence intervals, and when we look at the confidence intervals, we are confident that the rates are, indeed, comparable.

CHAIRPERSON ABRAMSON: Dr. Kweder.

DR. KWEDER: No.

CHAIRPERSON ABRAMSON: I'm sorry. We have a comment first over here.

DR. KWEDER: I'm sorry. Thank you.

CHAIRPERSON ABRAMSON: Are you okay?

Yes, Dr. Lewis.

DR. LEWIS: I just wanted to make another comment. We saw one slide where they actually looked at the UNOS liver transplant data for patients who underwent transplant or at least were listed for transplant for acute liver failure, and it always comes up, the issue of under reporting of events and, you know, we go round and round on this.

The most serious events always is under reported, is very minor events, and nobody thinks so, but no body can prove it, and I'm just wondering why we -- I mean, it ought to be fairly easy to do to look at the database on liver transplant patients, those who get a transplant and those who are listed but never get a transplant.

Now, that's not going to be everybody with liver failure because not everybody gets listed, but it would give us a much better idea if we want to look in this very specific area of drug induced hepatic failure, acute liver failure from drugs, whether it's all going to be acetaminophen or a few other drug as we've seen. It may give us a better handle. It would be a very important project, I think, to undertake, not just for this drug, as was done, but for all of the others because Will Lee's article and his acute liver failure group, which was mentioned here, in 17 centers, there's 110 transplant centers in the country. So obviously it's only a small fraction.

But it might give us a much better handle on some of these very important but rare events that, you know, we keep wanting to know what the signal is. Is it going to be one in 50,000?

I mean acute liver failure just spontaneously is one in a million in this country and probably higher in diabetics without drugs, and a number of other factors. But it's something that

could probably be done, you know, tangibly to get a better idea what's going on.

DR. WOODCOCK: I had -- I'm sorry.

CHAIRPERSON ABRAMSON: Dr. Woodcock, sure, of course.

DR. WOODCOCK: I had one other comment on behalf of the safety evaluator. Apparently some of the contact and investigation is still ongoing and so we do have additional -- there is some additional data other than what was sent to the consultants.

So we could straighten that out later. We just wanted to make that for the record. There's continuing efforts to investigate these cases, and some of that is reflected in the ODS consult.

DR. GOLDKIND: Right. That extra data is in the review. It wasn't in the initial reports.

CHAIRPERSON ABRAMSON: I think before we enter into a formal discussion on labeling, before we lose too many members, I think we can vote on Question No. 1. So why don't we do that?

Question No. 1 is: considering the

universe of available disease modifying therapies, is the benefit-to-risk profile for leflunomide acceptable for current indications?

And we've heard from Dr. Brandt and Dr. Williams that, yes, it was acceptable, and why don't we start with Dr. Gibofsky over here.

DR. GIBOFSKY: Yes.

DR. MANZI: Yes.

CHAIRPERSON ABRAMSON: I'm sorry. Yes.

MS. McBRAIR: Yes.

DR. ANDERSON: Yes.

DR. MAKUCH: Yes.

DR. ELASHOFF: Basically what we've seen for this drug seems to be reasonably consistent with what's seen for others.

(Laughter.)

CHAIRPERSON ABRAMSON: So that's a yes.

DR. FRIES: Yes.

And I also wanted to add my congratulations to the evolving signal monitoring of the AERS database because for the first time I actually think of it as an ongoing threat monitor

which can become more valuable with time and can go through a number of refinements, but I have always despaired of getting anything useful out of that data, and I think that we may be reaching a point in which we really can get some utility out of it. So I felt pretty good about that.

DR. DAY: Yes.

DR. LEWIS: Yes.

CHAIRPERSON ABRAMSON: Thank you.

So we have that recorded.

I guess the next question I think needs some discussion because the question is as we saw the data is there a signal coming. There clearly was something that came through in the adverse events that needed investigation. We saw a very comprehensive attempt to look at other databases.

And the question is: does the labeling need a modification because of the signal that came through with the serious adverse events, or conversely, is there enough data to support that signal?

And so I just want to open that up to

the committee. Is a label change warranted, at least as I read number two, based on the information that was seen?

I think I've posed the question. I'm curious what people might say.

Dr. Lewis?

DR. LEWIS: I think the label is satisfactory for all of the usual events that we talked about. The only question is, as I already mentioned, if there is a fatal case or a transplant case that is unequivocal, one case like that, I think, would warrant putting it in the label.

Again, even as we learn more about some of these cases, if you get the additional information, if it changes our vote from, you know, not enough data to possibly related or even possibility to probably related, again, we've already discussed that it's a risk-benefit decision that I don't think would change a lot.

We would obviously continue to look at signals like that. So for me, you know, it's going to be a decision for you to decide from the ongoing

database whether there's substantial information, maybe just one case that you would just add the words either "acute liver failure," which I think you could probably add. We all agreed that some of these cases were possibly related.

The question is: do you add anything more? Fatal, hepatitis, transplant, something like that?

And I think if you have it in the label, then I don't think it's going to detract from use. I think it's going to add one more layer perhaps to risk-benefit, but the benefit is still there.

CHAIRPERSON ABRAMSON: I guess the question that would come to mind as we saw in looking at the other data sets, that acute liver necrosis was not unique to leflunomide. So does that mean that each of these DMARDs should have a comparable kind of -- and from your perspective, Dr. Lewis?

I'm not suggesting that they should, but I'm just following the logic forward.

DR. LEWIS: Yeah. I think there's a

difference clinically between somebody who is in definite acute, you know, fulminant hepatic failure and needs a transplant or dies waiting for one; then someone who's just labeled as acute hepatic necrosis, whatever that means. I mean, that means the enzymes went up. Acute hepatic necrosis generally means you have a biopsy to look at or an autopsy or something, and you can get more information from it.

And we had very little of that information, you know, from the database that I looked at. So I think for me it would actually be the description of acute liver failure leading to transplant or death that's documented.

CHAIRPERSON ABRAMSON: Comments? Dr. Fries.

DR. FRIES: Yeah, I'd like to again raise this warning about the false positive signals. I'm interested in if other rheumatologists had the same experience.

When the Public Citizen memo became news and got on the front pages of papers, I had three

patients come and say, too, that they wanted to get off of leflunomide and one that said they didn't want to go on it because it caused serious liver damage.

Now, I don't think if you put that in the context of what Amy was telling us about her own experience that that makes sense, and I have this sort of gorge that rises when we have groups which are watchdogs for the public interest who may be hurting the health of the public by raising what turn out to be false positive red flags.

Now, I'm in favor of eternal vigilance, but until we actually have something that rises up out of background I don't think we ought to mention it.

CHAIRPERSON ABRAMSON: Other comments? Dr. Anderson.

DR. ANDERSON: We don't have the whole label to look at. The only part of it -- you know, in this context -- so the only part of it we have is actually from Dr. Day's presentation, and at the beginning of the warning section I don't know how

long the warning section is, but the whole paragraph here is included, which actually talks about elevations of liver enzymes already.

So there's already some mention of liver in there. So I don't know. I would agree with what Dr. Fries was saying. Until there's really a confirmed signal, you know, it's a false positive to do anything more than that.

DR. DAY: We've heard a lot about false positives. What about false negatives? Are any of us uncomfortable enough that it might be a false negative or the null hypothesis is sitting here?

CHAIRPERSON ABRAMSON: Well, my own sense is that we need more information, which is always an easy way out, but I would agree with Jim that we haven't seen compelling information from all of the other databases that there's a true signal there, and so to put something in the label when you could probably find other drug reports for other drugs that you then would then have to go back and put in their label, I'm not sure the evidence bears it out personally.

I think that Dr. Seeff raised another question, which is when you look at the labels, there's the issue of how long do you continue to monitor, and what does it really mean to have chronic twofold elevations of AST or ALT.

And I think that's an area that we need more information on, but to put, in essence, anecdotal reports into the label without firm confirmation is of some concern for me based on the information that we've seen.

DR. FAICH: I just wanted to add one thing. This issue of a false negative maybe should be addressed.

I'm Gerry Faich. I'm an epidemiologist.

I would just like to point out to the committee that the sum total of patients studied in a controlled environment, meaning the clinical trials, plus the Aetna study, plus the Protocare study, plus the national database amounts to well over 20,000 patient years.

Within that, there are three possible cases, one in the trial that was the only elevated

liver enzyme case that you heard about which reversed; one labeled hepatic necrosis in the Aetna study; and one case that was in the national database that was associated with sepsis.DR. MANZI:

So the numerator at best is three in settings where it's highly likely that all cases were captured. it is also clear that those three cases may have been confounded, may be related to the underlying disease, may have been related to methotrexate. All of those are possible.

But the point is it seems to me once you have a signal for spontaneous reports, what you want to do is do good epidemiology in sizable populations. That's been done here. I would submit that that data is strong enough to suggest that there is -- I don't think it absolutely rules out a risk, but it very strongly points in the direction that if the risk is present at all, it must be very small.

CHAIRPERSON ABRAMSON: Yes. Oh, I'm sorry. Yeah.

DR. ELASHOFF: I just wanted to comment

that although this is extremely difficult from a statistical point of view or from the point of view of estimating things, that the issue of what our best estimate of a rate is and what rate would be too high under the circumstances, you can't even sort of say how many patients you need to study or how big the thing needs to be until you have some notion of what rate is too high a rate.

And that also applies to the issue of labeling, and one person said if there's one confirmed case, he thinks that that should warrant labeling, but perhaps we should give some thought over the future to what rates are important enough in any given context that we think that they need to be reported.

At some level, somebody who's going to get any given drug is going to have almost anything happen to them because everybody dies in the end anyway, and so that it seems to me we need to give some thought to what rates are common, what rates are of concern, what rates are ones that ought to trigger a label.

And I know this is an extremely difficult thing to think about, but I think it would be of some use to discuss things in that way.

CHAIRPERSON ABRAMSON: Dr. Gibofsky.

DR. GIBOFSKY: I'm swayed by one of the comments that Dr. Strand made pointing out to us that this is a bad disease currently with limited therapeutic options. It's important to realize that the other agents available to us do not work on 100 percent of patients. Our ACR 20s are acceptable, but they're not desirable. Even our 50s and 70s are not that.

And I think at the end of the day we're aware of the risk of these agents and we enter into the appropriate dialogue with our patient as to what the risks are versus the benefits.

As I tried to tease out of Jim Freeze earlier, when you look at the domain of the five Ds, how do the patients weigh things?

And clearly there are patients who will say, "I would rather spend more and be less disabled." "I would rather be more disabled and

spend less."

We make those tradeoffs, but I think to the extent that we can make our patients aware that nothing is without risk, this is not without risk. Nothing that we are going to attempt to use is without risk, but we're going to watch you, and we're comfortable managing the risks. I think it's a risk that ought to be take, particularly when our patients are individual in their responsiveness to therapies and our patients do not respond acceptably -- all of them do not -- to the other therapies.

It was suggested that perhaps one could practice medicine or rheumatology without this drug. Sir William Osler practiced medicine without penicillin. I'm not sure I would want to, and I think this is an acceptable alternative to the current medications that we have for those patients who respond to it.

CHAIRPERSON ABRAMSON: All right, and I think that's an important point also, that even these biological drugs that have really changed the way we think about RA and ACR 50 response, 50

percent or fewer of the patients. So that leaves an awful lot of the people who need alternatives. We haven't really stopped this disease, and I think that's a common misperception perhaps, that the drugs are so effective that we don't need others.

So I guess to the FDA, are the comments about the label sufficient or do you want something more specific from the panel.

DR. SIMON: I only wonder whether or not the panel thinks that we need -- because we do believe that the labeling needs to be changed slightly -- that there needs to be a little bit more emphasis to potential liver toxicity. One wonders whether or not we need to do any other kind of risk communication, "Dear Doctor" letter, letters and information from us as the FDA.

What does the panel feel about that?

DR. LEWIS: If you want me to start, I would say no. If you change the label and you put in one more layer of liver toxicity, it's already pretty well replete with things that happen in the liver. If you're going to go to, you know, one case

of liver failure has been reported or whatever or one transplant has occurred, I don't think that rises to the level that I would want to see a "Dear Doctor" letter or anything else about that.

I mean, if you accumulate additional information, that's different, but on the basis of what we've discussed today, I don't think it would be necessary.

CHAIRPERSON ABRAMSON: The only comment I would add though is that it's important that information be communicated to the other side, that what we heard today is that these reports are there, but review of multiple databases, or however one would frame it, does not seem to indicate an enhanced risk to this drug.

So we're reporting this. We need more information, but the hazard is the one that Jim keeps coming back to frighten people about something that we're still not certain about.

MS. McBRAIR: As a patient educator, I think some of the changes to the label that Dr. Day suggested will give greater emphasis to the concerns

that people have about the drug and about how physicians manage it and work with their patients. I think that will be wonderful just in itself.

CHAIRPERSON ABRAMSON: Other comments?

(No response.)

CHAIRPERSON ABRAMSON: Well, with that, I guess I would thank everybody and turn it back to Dr. Simon.

DR. SIMON: Well, first I think that you have educated us significantly about this particular problem. We are incredibly grateful. We recognize that the amount of information and the time it took to prepare yourselves for this particular meeting was quite onerous, and again, we thank you for making yourselves available to give us such cogent information, and I congratulate the chair on running such an incredibly efficient meeting even without the break.

So thank you very much.

(Whereupon, at 4:52 p.m., the meeting in the above-entitled matter was concluded.)