TRANSCRIPT OF PROCEEDINGS

DEPARTMENT OF HEALTH AND HUMAN SERVICES

FOOD AND DRUG ADMINISTRATION

CENTER FOR DRUG EVALUATION AND RESEARCH

ARTHRITIS ADVISORY COMMITTEE

DESIGN AND ASSESSMENT OF CLINICAL TRIALS OF DRUGS,

BIOLOGICS AND DEVICES THAT ARE BEING DEVELOPED

FOR TREATMENT OF SYSTEMIC LUPUS ERYTHEMATOSUS

Pages 1 thru 259

Silver Spring, Maryland
February 23, 1999

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

DEPARTMENT OF HEALTH AND HUMAN SERVICES

FOOD AND DRUG ADMINISTRATION

CENTER FOR DRUG EVALUATION AND RESEARCH

ARTHRITIS ADVISORY COMMITTEE

DESIGN AND ASSESSMENT OF CLINICAL TRIALS OF DRUGS, BIOLOGICS AND DEVICES THAT ARE BEING DEVELOPED FOR TREATMENT OF SYSTEMIC LUPUS ERYTHEMATOSUS

This transcript has not been edited or corrected, but appears as received from the commercial transcribing service; the Food and Drug Administration makes no representation as to its accuracy.

Tuesday, February 23, 1999

8:15 a.m.

Holiday Inn Silver Spring
8777 Georgia Avenue
Kennedy Ballroom
Silver Spring, Maryland

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

PARTICIPANTS

Steven B. Abramson, M.D., Acting Chairperson
Kathleen Reedy, Executive Secretary

MEMBERS

Nigel E. Harris, M.D.
Daniel J. Lovell, J.D., M.P.H.
Leona M. Malone (Consumer Representative)
Frank Pucino, Jr., Pharm.D.

FDA CONSULTANTS

Kenneth D. Brandt, M.D.
Leigh F. Callahan, Ph.D.
David T. Felson, M.D., M.P.H.
Felix Fernandez-Madrid, M.D., Ph.D.
Ildy M. Katona, M.D., CAPT, MC, USN
Matthew H. Liang, M.D., M.P.H.
Harvinder S. Luthra, M.D.
Kevin R. McConnell, M.D.
Larry W. Moreland, M.D.
Michelle A. Petri, M.D., M.P.H.
Yvonne Sherrer, M.D.
Earl D. Silverman, M.D.,FRCPC
Barbara White, M.D.

GUEST EXPERTS

James E. Balow, M.D.
James V. Donadio, M.D.
Paul Fortin, M.D., M.P.H., FRCP(C)
Ellen Ginzler, M.D., M.P.H.
Ken Kalunian, M.D.

GUEST SPEAKERS

Dimitrios T. Boumpas, M.D.
Dafna D. Gladman, M.D., FRCPC
David Isenberg, M.D., FRCP
Edmund J. Lewis, M.D.
Vibeke Strand, M.D.
John H. Klippel, M.D.
Sterling West, M.D.

FDA

Robert Delap, M.D., Ph.D.
William D. Schwieterman, M.D.
Jeffrey N. Siegel, M.D.

C O N T E N T S

C O N T E N T S (continued)

P R O C E E D I N G S

Call to Order, Introductions

DR. ABRAMSON: Good morning. I am Dr. Abramson.

We would like to begin this morning, prior to the presentations, to ask people on the panel to please introduce themselves.

Dr. DeLap, would you like to begin?

DR. DeLAP: Robert DeLap, Center for Drugs, FDA.

DR. SCHWIETERMAN: Bill Schwieterman, Clinical Trials, CBER, FDA.

DR. LUTHRA: Harvi Luthra, rheumatologist from Rochester, Minnesota.

DR. WHITE: Barbara White, rheumatologist, University of Maryland.

DR. FELSON: David Felson, rheumatologist from Boston University.

DR. FERNANDEZ-MADRID: Felix Fernandez-Madrid, rheumatologist from Wayne State University, Detroit.

DR. LOVELL: Dan Lovell, University of Cincinnati.

DR. SHERRER: Yvonne Sherrer, rheumatologist, Fort Lauderdale.

DR. McCONNELL: Kevin McConnell, nephrologist, Charlottesville, University of Virginia.

DR. PUCINO: Frank Pucino, pharmacy, National Institutes of Health.

DR. LIANG: Matt Liang, rheumatologist, Harvard Medical School.

DR. CALLAHAN: Leigh Callahan, University of North Carolina, Chapel Hill.

DR. MORELAND: Larry Moreland, rheumatologist, University of Alabama.

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

DR. ABRAMSON: Steve Abramson, rheumatologist, Hospital for Joint Diseases.

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

MS. MALONE: Leona Malone, Consumer Representative.

DR. HARRIS: Nigel Harris, rheumatologist, Dean, Morehouse School of Medicine.

DR. SILVERMAN: Earl Silverman, rheumatologist, Hospital for Sick Children, University of Toronto.

DR. KALUNIAN: Ken Kalunian, rheumatologist, UCLA.

DR. GINZLER: Ellen Ginzler, rheumatologist, SUNY, Health Science Center at Brooklyn.

DR. FORTIN: Paul Fortin, rheumatologist, McGill University, Montreal.

DR. STRAND: Vibeke Strand, rheumatologist, Clinical Faculty, Stanford.

DR. GLADMAN: Dafna Gladman, rheumatologist, University of Toronto.

DR. PETRI: Michelle Petri, rheumatologist, Johns Hopkins.

DR. ISENBERG: David Isenberg, rheumatologist, University College, London.

DR. ABRAMSON: Thank you. Clearly, we have all of the rheumatologists in this country around this table.

We next would like to have a meeting statement by Kathleen Reedy.

Meeting Statement

MS. REEDY: 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 since the issues to be discussed by the committee will not have a unique impact on any particular firm or product, but rather may have widespread implications to all similar products in accordance with 18 United States Code 208(b)(3), general matters waivers have been granted for today's meeting.

In the event that the discussions involve any 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.

DR. ABRAMSON: Thank you very much.

We will now move to the introductory comments.

Dr. DeLap, please.

Introductory Comments

DR. DeLAP: I would just like to say welcome and I really appreciate our distinguished panelists taking time from their busy schedules to be here and to help us in thinking about some of the issues on clinical trials in lupus, and I would like to thank our guests for also attending. I look forward to interesting and informative discussions today.

DR. ABRAMSON: Thank you.

Dr. Schwieterman.

DR. SCHWIETERMAN: I have only five brief slides that I mean to provide as simply a brief perspective and orientation for this particular meeting.

[Slide.]

As I was driving in to work here today, I realized that I had omitted the most obvious perspective and orientation, and that is, that the field of rheumatology is rapidly changing toward many agents that are directed including some of the biological agents of particular elements of the immune system, and because of these new advances in therapeutics, many of which are in biologics, some of which are in the Center for Drugs, we are at indeed a threshold at which we can begin thinking about larger clinical trials for this particular disease and all the details that are included therein.

[Slide.]

The five points I want to make are simple ones, but I think they bear mentioning. We definitely are at the beginning of understanding how best to do clinical studies in SLE. I did a literature search myself late last night, looking to see just the relative numbers of clinical trials in some different diseases.

These are controlled, randomized clinical studies. It is probably not surprising, but, nevertheless, it is a little bit informative to find that there are only nine references to any clinical studies in lupus. This is any sort of study to do with SLE, 52 in rheumatoid arthritis and 277 in diabetes.

I think these numbers are going to change in the

near future as we become more sophisticated about how to study this disease's biological therapies and the other drug therapies evolve. I think this is an exciting time to be thinking about how to perform and how to improve the performance of these agents.

[Slide.]

Second is an obvious point to almost everyone in the room. SLE is a difficult disease to treat and a difficult one to study. For a number of reasons I have picked three here. There are different disease manifestations and subpopulations. It's a chronic waxing and waning disease. Surrogate markers can be difficult to interpret as indicators of clinical efficacy.

I think this is something to keep in the back of your mind as we have the discussions today because in all likelihood, we are going to get a variety of opinions about how best to use endpoints, how best to use inclusion criteria, how best to orient the trials toward a particular endpoint, and I don't think that the field is particularly acrimonious, I just think that the disease is particularly difficult, and I think we have to keep this in mind as we talk about all the different details.

Thirdly, Vibeke is going to talk about this later today, but I just did want to mention it in brief mention here. The OMERACT conclusions from last year provided a

framework for today's discussion. Many rheumatologists from around the world got together, decided on four particular parameters that ought to be included in any study of SLE including the following four: disease activity, health-related quality of life, damage assessment, and toxicity and adverse events.

From here, I think we can use this as a springboard into the more detailed discussions of how best to measure disease activity, how best to measure quality of life, damage, and so forth. I think this is a good start, and I think today can be a progression on that.

[Slide.]

Fourthly, good guidance practices are in effect today. The FDA has thought about this, but we are keenly interested in people's opinions here simply because we need a starting point by which to put together a straw man that we can bring back to this committee, with which to then write a definitive guidance document.

It can be a long arduous process, and we have no illusions about it being an easy one, but we have every intention, following this particular meeting, of sitting down and discussing what people thought or think about the broad issues toward improving product development including things like whether claims are appropriate for a guidance document, what are the best endpoints, the types of

analyses, the types of patient populations that ought to be studied, and so forth.

[Slide.]

Last but not least, I wanted to end on an optimistic note because I feel optimistic about this. The rheumatologists have in fact, in my experience anyway, been quite successful in tackling the difficult clinical trial design issues. It is a long history of fruitful meetings spanning over the past two decades.

I think that the ACR criteria in RA in 1993 greatly facilitated the development of that field. The guidance document that has been written by the FDA has not only been reasonably well received, I think quite well received in many respects, but also it has been used as an example for other guidance documents within the agency with regard to the kinds of things, the kinds of ideas that were first put into that, claims, and so forth.

Finally, I know that there are a number of other similar endeavors in other areas.

So, with that as an orientation, I very much look forward to this meeting, and look forward, then, to taking the results of this meeting forward even further.

DR. ABRAMSON: Thank you very much, Bill.

Next, Dr. Siegel.

[Slide.]

DR. SIEGEL: Welcome to the special Advisory Committee Public Workshop on Clinical Trials in Lupus.

[Slide.]

The effort that we are undertaking today is part of a general effort at the FDA that has been undertaken over the last few years to develop guidance documents for the clinical development of new agents for a study in rheumatic diseases. These efforts are intended to provide guidance to industry about their clinical development programs on these various indications.

The first document which was worked on was rheumatoid arthritis, and this was completed a while ago and has now been published in its final form.

The agency then began on developing a guidance document on osteoarthritis, and there have been several meetings on this, and this guidance document is in progress.

The third area that we are beginning to undertake is lupus, and that is the purpose of this meeting. This meeting is really the initial stage of developing a guidance document for lupus.

We begin, as Bill Schwieterman mentioned, the process of developing a guidance document by seeking community input about how new agents should be developed for lupus, and then it will go through a series of other steps before a final document with several times available for

public comment.

[Slide.]

Some of the issues that the FDA is seeking input from the community on with regard to clinical trials in lupus are shown on this slide.

The first one, we would like to learn what the community feels are responses which represent a meaningful clinical improvement in lupus, and this would presumably be the basis for determining what claims could be recognized for new agents as studied in lupus.

Second, we are very interested in finding out what the community feels are appropriate assessments which could be used as a primary efficacy endpoint in the clinical trial.

Let me just talk a little bit about that. I think some people feel it should be straightforward, you just measure many different things which cover all of the important areas in lupus, and then look at the results of a trial, but it is somewhat more complicated than that, because the more endpoints you study, the higher likelihood there is that any one will show a positive result by chance alone.

So, a common practice in clinical trials conducted by industry sponsors is to choose a single primary endpoint which determines the success or failure of a clinical trial.

So, we are very interested in hearing what the group feels could be used as primary endpoints for clinical trials.

The third area that we are interested particularly in input on is what the role of surrogate markers of efficacy should be in lupus. I think the best way to explain this is that if you can use a clinical marker to determine efficacy, that is usually the best way to conduct a clinical trial, but in many cases, the clinical endpoint takes many years before it can be assessed, and it would be impractical to wait five or 10 years to determine the benefit, and if there are reliable surrogate markers, laboratory markers or otherwise, which are highly associated with, and predictive of, a beneficial clinical outcome, these can used as clinical markers of efficacy, and we are very interested in certain areas in determining what you all feel could represent adequate surrogate markers.

This is particularly true for lupus nephritis where many of the clinical endpoints, such as mortality and progression to end-stage renal disease and dialysis, may take many years, but there are other potential markers that can be used as surrogates.

[Slide.]

I just wanted to talk for a minute about an example of a claim structure, and I have chosen the claim structure that was used in rheumatoid arthritis as the

example here.

This is the claim structure that was decided upon as a result of many public meetings, as well as internal discussions within the FDA, and it is the basis for a guidance document that was just published.

It was felt that a single claim was not adequate to capture the full range of disease manifestations in rheumatoid arthritis, and instead, five separate claims were decided on as the basis for determination of efficacy.

The first is a signs and symptoms claim, which is based on a clinical trial of at least six months in duration, showing a benefit in signs and symptoms, often based on a validated index, such as the ACR 20.

The reason why six months was chosen -- and I go into this because this may be something that you all will want to consider -- the reason that six months was decided upon was that we wanted to make sure that we had evidence regarding the durability of a clinical response, that it wasn't just transient, and also to determine the long-term safety and short-term trials are not adequate often to determine the long-term safety.

The second claim is prevention of structural damage, and the idea here was that signs and symptoms alone didn't capture the tendency of rheumatoid arthritis to lead to joint destruction over time and disability. So,

prevention of structural damage usually based on x-ray endpoints was considered as the second claim.

The third claim is a major clinical response intended to go well beyond the modest benefit that had been seen with some previously available agents, and it was decided that while an ACR 20 percent response was sufficient was a signs and symptoms claim, it was decided that a 70 percent response for six consecutive months would be required to demonstrate a major clinical response.

Next was a complete clinical response based on absence of disease for six consecutive months or remission, the same thing without requirement for antirheumatic agents, and finally, a durability claim, the last one, entitled "Improved Physical Function Disability," based on improved function in the two- to five-year trial.

Now, that is rheumatoid arthritis.

[Slide.]

I think some of the challenges that we face in trying to devise clinical development guidelines for lupus, some of the challenges are shown here.

Unlike rheumatoid arthritis which attacks primarily the joints, there is marked heterogeneity of which organ systems are attacked in lupus.

Second, there is no single index which has the broad acceptability of the ACR 20, although there are a

number of candidates which are highly promising.

Third, there is generally a paucity of randomized controlled interventional trials showing efficacy to be used to validate endpoints as being sensitive and specific.

[Slide.]

In contrast, there have been a number of recent advances over the last several years, which will make our job somewhat easier.

First, there has been tremendous progress in the measurement of disease activity both by devising reliable and valid disease activity in disease which allow patients with varying manifestations of disease to be compared, and also there has been a lot of work in defining flares of lupus.

Second, there has been a lot of advances in the area of pathophysiology of disease and the genetics of human disease, as well as animal models, giving insights into human disease, and I think we stand at the point now where there are a number of promising therapeutic agents which are currently in clinical trials, which will hopefully be the basis for improved therapies in the future.

[Slide.]

Now, I want to say that we know that there are many controversies in the field right now, and we do not expect to develop a consensus in many areas. Rather, we are

interested in hearing the feeling at the present time about what would be an adequate basis for conducting a clinical trial.

Some of the potential outcomes of this meeting and of devising a guidance are shown here. First, we hope to define claims which would cover the full range of potential clinical benefits in lupus. As a result of devising some kind of structure for clinical development, we hope this will provide incentives for the development of effective therapies in the future.

Finally, in those areas where there is a lot of uncertainty, we hope that this work will define areas in need of further research.

With that, I will stop and listen to the presentations.

DR. ABRAMSON: Thank you very much Dr. Siegel.

The way the day is divided, as you can see from the agenda, is that in the morning we will try to grapple with lupus as the broad disease, the heterogeneous disease, and hear from individuals who have spent a great deal of time thinking about this disease and developing disease activity indices and the definitions of flare and damage indices.

We will have thereafter a panel discussion to open these presentations up to discussion.

In the afternoon, we will deal more directly with specific end organ damage particularly lupus nephritis.

I think this morning is a particularly arduous task to try and hear how people have tried to develop indices and what we will do is we will ask each presenter to make their presentation. We will hold questions until the period of the panel discussion unless there are specific questions of clarity or specific to the presentation.

With that, I would like to introduce Dr. Isenberg from the University College of London to discuss disease activity and health status.

Efficacy Assessment for Clinical Trials in SLE

Disease Activity: Health Status

DR. ISENBERG: Let me start by thanking you for the pleasure and the privilege of being here. I was interested to see downstairs in the lobby that there is a big notice which says, Welcome to the American Hypnosis Society, and I thought if anybody goes to sleep in the next 20 minutes, that would be good because it could mean I could have another job.

[Slide.]

The guts of what I want to tell you about this morning is really on this first slide, and that is, in order to understand the totality of the effect of a disease like lupus upon a patient, I am a firm believer that we have to

have measures of disease activity, by which I mean clinical features which can be corrected or improved; damage, by which I mean permanent change; and patient perception.

The importance of patient perception was brought out to me very forcefully some years ago when we asked 100 of our patients in London to name the one feature of their disease which troubled them the most.

In my naivete I had assume that the answer would be most likely fear of facial disfigurement or death or renal failure, and the answer was fatigue. That was the thing which troubled the patients the most.

So, it is terribly important I think that we have to have good parameters for measuring all three of these concepts.

[Slide.]

As Matthew Liang has pointed out, between the middle fifties and the middle eighties, some 60 different global score activity indices were produced by different groups around the world, and they were all unsatisfactory because they were unreliable, they had never been validated, and I feel as guilty as anybody else having contributed at least two of those disease global score indices to the literature, and it was really for the reasons which we have already heard about, the remarkable heterogeneity of lupus, that in the UK some 15 years ago, a group of rheumatologists

began to meet, and we continue to meet to this day every three or four months, to try to devise a rather better way of looking at disease activity.

[Slide.]

We thought in particular we wanted to come up with a better scheme which could reflect the remarkably heterogeneous nature of the disease, and in order to do that we introduced what is called the BILAG system, and this divides lupus activity into eight different systems as indicated on this slide - the general features, the mucocutaneous features, CNS, musculoskeletal, cardiovascular, respiratory, vasculitis, renal, and hematological, and we based it on what we called the "physician's intention to treat." In other words, we, as a group of clinicians, got together, went through all of the features that we felt were due to lupus, and found that there was a considerable measure of agreement as to what we thought we would do were a patient to develop that particular problem.

The system is computerized, so that once the data has been given to the machine, the computer will automatically work the score out for you, and I am going to show you a bit of that information later on.

[Slide.]

Now, within each of these eight organs or systems,

we have a number of clinical questions which the patients and we answer, and the disease activity for each of these eights organs or systems is divided into A for action, meaning that major immunosuppressive therapy is required; B for beware, meaning that the patient is already known to be active, and no major change in treatment is required; down to C for contentment, low level activity only; D for discount in the sense that the disease in this system or organ is not completely resolved; and E for no evidence of any disease in the system now or previously.

So, this applies to each of the eight organs or systems.

[Slide.]

Now, this is just one single example of how this works. So, a patient comes into the clinic and we inquire whether there is any evidence of pericardial pain, dyspnea, evidence of cardiac failure, and so forth, through this list.

The clinician fills in "1" meaning the symptom is improving, "2" meaning that it is the same as the last visit, "3" that it is worse, "4" that it is new. In some cases, we simply require yes/no answers. This is for cardiovascular/respiratory system.

[Slide.]

Now, in order to get a Category A in this

particular system, the patient has to have either cardiac failure or symptomatic effusion plus two other criteria as listed on this slide.

Now, I have with me one of the main papers that we produced, which describes this index in great detail, and if anybody has not seen this, I will be very happy to give you the full paper subsequently.

Category B will be any two of the criteria listed under A, but in the absence of cardiac failure or symptomatic effusion, whereas, Category C in this group will be mild, intermittent chest pain or just one of those other features, D for previous involvement but no evidence of activity currently, and E for no evidence of previous involvement.

The same applies to each of these other seven organs or systems.

[Slide.]

So, in practice, what actually happens is the patient comes to the clinic, a patient assessment form, which I also have with me today if anyone wants to see it, can be completed manually or, if we have a laptop, it can be inserted into the computer. In the current system, the IBM or Apple Mackintosh versions are available, but we are just switching over to a new system that I am going to tell you about in a moment.

Once the information has been provided to the computer, the hematology scores have to be added later, the patient can be identified on an A to E category in eight of these eight organs or systems.

The beauty about it is that the clinical score can then be matched to a stored serum sample if that is what you want to do with it.

[Slide.]

In practice, this is the sort of thing that we can see. So, here is one patient, a real live patient, and here are some dates during which she was seen, and you can just see just by glancing at this, you get a very good idea. You can see not only when this patient became active, but in which particular systems. If you want to use the word "flare," you can see that this patient flared exactly seven years ago. She went to a Grade A from a C, and the general features, and she went to a B from a D, mucocutaneous.

She developed serious arthritis, as you can see, and there was also evidence of renal disease for the first time.

I think you can see that this would lend itself very nicely to a drug study, and I will show you how we are doing that a little bit later on.

[Slide.]

Now, clearly, you wanted to validate this system,

and one of the beauties of the systems is that it is a testable hypothesis. So, here what we have done is to look at the patients who actually developed a BILAG Grade A to see if that actually meant that they had disease-modifying treatment, were they actually given large doses of steroids after they developed the Grade A.

We got an independent observer to go back through the hospital notes of several different hospitals, several of the lupus clinics in the UK, and you can see that in five out of six cases, an A in the general system was indeed followed by disease-modifying treatment, in 24 out of the 29 cases judged to be Grade A, mucocutaneous, and so forth.

You can see that in general terms, with the exception of the CNS, which I still think remains the most difficult aspect of lupus to assess, you can see that the system stands up pretty well to examination.

[Slide.]

We wanted to see how reliable it was, and in order to do this, we got an independent observer to go around to five different lupus clinics in the UK, and this naive observer was asked to examine the patient after the patient had been examined by the local rheumatologist, and to work out the score.

You can see that there was general agreement. This was on two occasions, the first assessment and the

subsequent assessment, and you can see there was a considerable measure of agreement down each of these clinical features, each of these organs or systems.

[Slide.]

As I said, one of the beauties of the system is that it enables us to look at stored serum samples. We can use this BILAG system to look for disease activity and to see whether this correlates with any particular organ or system.

A colleague of mine, Michael Ehrenstein, as an example here, has identified an idiotype on a monoclonal antibody to double-stranded DNA B3id, so the B3id is over here, and he wanted to determine whether or not this idiotype was associated with any particular aspect of lupus or indeed whether it was present in any disease control.

You can see that what raised levels of this idiotype do is to pick out the patients with activity in the musculoskeletal system, but really very little evidence of it being raised in patients with renal lupus, cardiovascular/respiratory lupus, mucocutaneous lupus, or indeed any of the disease controls that we looked at, Sjogren's, rheumatoid, or myositis.

We have also looked in serial bleeds in a study with Mericin, asking the question whether any eight or nine different antibodies that we were looking at, whether they

were correlated with disease activity in any of the eight organs or systems, and again the answers are pretty clear-cut, antibodies to double-stranded DNA strikingly associated with renal disease, to a considerable extent with cardiovascular/respiratory disease, much weaker associations of DNA with global score, and antibodies to ribosomal P with musculoskeletal involvement or hematological involvement, and with general features, but relatively few good correlations were found, in fact, in this study.

[Slide.]

Now, we were hardly alone in wanting to try to improve in a sense the lot of lupus research and the lot of the lupus patients. This picture was taken 12 years ago when, under auspices of NATO, a number of groups with an interest in lupus activity got together, and some of you will recognize Matt Liang over here, Dr. Gladman over there, and myself over there.

[Slide.]

The purpose of this meeting was to try to agree a set of principles. We wanted to explore various lupus activity indices that we felt were worthwhile. We wanted to look in detail at the BILAG and compare it to the SLEDAI system, the global score system, and the SLAM global score system, which I am going to tell you about in a moment.

We wanted to consider the possibility of setting

up a database, we wanted to consider deriving a damage index, and we wanted to consider deriving a patient health assessment questionnaire. This was back in 1987, well before the OMERACT ideas came forth and really confirmed that these were worthwhile pursuits.

[Slide.]

A little bit about the SLEDAI index, which was developed by Dr. Gladman and Murray Urowitz, and their colleagues, and this index was constructed following studies in which clinicians rated the importance of 37 variables which had been preselected as likely to define lupus activity. An analysis by this group showed that 24 variables were the most important and appeared to be contributing majorly to the clinical judgment of activity.

[Slide.]

So, the SLEDAI system includes these 24 descriptors in nine organs or systems with weighting being assigned that is based on multiple regression techniques. Each item is considered to be present or absent over 10 days before the assessment, and obviously, the hematology and immunology results were found to be of little benefit, and that was the same as largely true in the BILAG system where we did not find many autoantibodies to be very helpful to us, and they form only a minor part of this index.

[Slide.]

At much the same time, Matt Liang and colleagues were developing the SLAM index. This uses disease manifestations which had been culled from the literature and have been refined by a group of interested clinicians. The items were chosen on the basis of those which could be graded in preference to those that could not be graded, and those which could be operationally defined and reliably tested in preference to those which could not be.

[Slide.]

They devised a scale which includes these 24 clinical manifestations and 8 laboratory tests which evaluated the organs which cannot be assessed otherwise. The hematology system is the classic example.

The scale refers to the month prior to the assessment, as indeed does the BILAG system.

[Slide.]

Two dimensions are incorporated into the SLAM if a manifestation is active or not, and then severity is then used to expand the scale judged by the need to treat with immunosuppressives, the need to follow the patient more closely, and the functional or prognostic consequences of the manifestation.

[Slide.]

Now, we were very anxious to compare these indices, and in order to do that, we had to convert the

BILAG into a global score although as I have said, this was not its fundamental intention.

The way that we did that was to ascribe 9 points to an A score, 3 points to a B, 1 point to a C, and no points to a D or an E.

There have been a number of studies that have been published by this group, which has worked very collaboratively together. Again, I have some publications with me, if anybody wants to, I can get them afterwards.

This is just one study that was undertaken by the Toronto group in collaboration with physicians from around the world, in which we examined seven patients with lupus, seven different observers from I think five different countries.

Seven centers were brought together in Toronto, representing a spectrum of lupus manifestations and activity. They were each examined by up to four of a panel of seven observers using the Youden square design, and each observer completed all three indices and a category rating scale for disease activity.

The results were very satisfying in that the p-values were all highly significant and, as I said, we went on to undertake a number of other studies including studies of sensitivity to change, for example.

[Slide.]

Subsequently, some of you I am sure will know this, our colleagues in Italy have developed a so-called ECLAM scoring system, which is also a global score system, and Stefano Bombardiere and colleagues then performed a study in which they compared five scoring systems - the SLAM I have told you about, the SLEDAI, and the BILAG, the ECLAM and the SIS system are two other global score systems. In a study of 75 patients, they showed a remarkably good level of correlation between each of these different indices.

[Slide.]

I want to turn now for a few moments to the question of the SF36 and patient assessment. Nobody as far as I am aware has devised a health perception index purely for lupus, but it does appear, as I will show you, that the SF36 can provide us with such an instrument.

This was developed by John Ware and his colleagues initially at the Rand Corporation, subsequently in Boston. They developed a series of questionnaires designed to measure health attributes using multi-item scales.

In the original questionnaire there were over 200 questions. They subtracted these down to what is called the SF20 or the Short Form 20, and this has been used in a number of studies, as I will show you in a moment.

However, it was felt that the SF20 was failing to capture certain important problems in lupus patients, and in

particular, the question of fatigue is not directly tackled in SF20, so a more extensive questionnaire, the SF36, has been devised.

It represents a compromise between what is regarded as user-friendly and what is regarded as sufficiently detailed.

[Slide.]

The SF36, therefore, tries to identify problems of physical functioning, role limitations due to physical problems, emotional problems, social functioning, mental health, general health perception, and general health change. It also incorporates energy/fatigue, and fatigue in separate scales.

[Slide.]

Now, we were interested in comparing the SF20-plus which is simply the SF20 with one extra question on fatigue and the SF36, and in collaboration with colleagues in Birmingham, England, we have looked at 150 lupus patients and have shown that there is a very strong correlation in all those matching areas of SF20 and SF36.

So, whether we are using the SF20-plus or the SF36, it does appear to provide a useful measure of health assessment in lupus patients.

[Slide.]

This is just to show you very briefly a study of

141 patients in which we have looked at the damage index, which we are going to hear about a little bit more from Dafna Gladman in a moment, and health status, and you can see that as might have been predicted, there is relatively little correlation between the SF20-plus, the patient perception index, and aspects of damage, which is shown up here in the yellow, the purple indicates that there is correlation.

[Slide.]

There are more correlations between the BILAG disease activity and the SF36 in particular. The general features under the BILAG system correlate with a number of health perception problems. Again, that might have been predicted.

[Slide.]

Where we see correlations between damage and activity, those tend to be in areas which make sense, so disease activity in lupus does show correlation with the damage in the musculoskeletal system. Likewise, activity and cardiovascular/respiratory disease associated with damage in the cardiovascular/respiratory system.

[Slide.]

Finally, just to update you on a couple of ongoing things just to let you know what is going on at the moment. One of the criticisms that the BILAG group has tried to

correct is that we have not previously and up until recently used the BILAG system in a clinical trial. That is what we are now doing at the moment.

We are comparing cyclosporin against azathioprine in patients with active disease. So, the aims are to compare the effectiveness of Neoral versus azathioprine in patients with severe lupus, and we look for steroid-sparing effects as a primary endpoint, and we are also looking at the number of flares, which we defined as a new BILAG, Category A, or a B score coming from a D or an E.

We look at immunological outcomes, we look at toxicity, we look at patient perception using SF36, and we are using the damage score, and that is an ongoing study at the moment.

[Slide.]

Also, something which I think may be useful for some of the lupologists here in the audience is that we have just developed a new computer system. What this system does or what it provides for us is a considerable amount of basic demographic information. It incorporates the year in which individual ACR criteria for lupus were met and thus the date of diagnosis.

It gives us an activity index every time the patient visits the clinic. It provides full details of the therapy at each clinic visit. It records laboratory results

at each clinic visit. It records the damage index, which can be done every 6 to 12 months. It records a patient perception index at 6 to 12 months. It also provides a notable graphing capacity, so if we wanted to compare, for example, C3 against a particular item of damage, we can do it. If we want to compare DNA antibodies against a certain measure of disease activity, we can do that also.

[Slide.]

Finally, just to confirm that in my view we now have available a number of validated good, useful global score systems including SLAMs, SLEDAI, SIS, ECLAM, and the LAI score, but if you want something a little bit more sophisticated, I don't think the BILAG score can give you that.

[Slide.]

But more than anything, I do want to try or hope I have convinced you that in order to, as I say, to study the effects, the totality of effects of lupus on a patient, we really have to have disease activity scores, disease damage scores, and patient perception index.

Thank you very much.

DR. ABRAMSON: Thank you, David.

We have a couple of minutes if people have questions for clarification of Dr. Isenberg.

David.

DR. FELSON: I have a bunch of questions for you, David. First, it looked like the BILAG was a transitional question-based instrument meaning that for each of those organ systems you asked whether there was increased involvement, static involvement, decreased involvement. It isn't a state, it is a transition-based instrument.

Is that true of the other instruments also?

DR. ISENBERG: It is probably more appropriate to actually ask the guys who actually developed the instrument including Matt and Dafna what they feel about that. You are right, that is the way that the BILAG was constructed.

DR. FELSON: So, conceivably, one could have extraordinarily active disease and have an unchanged or even an improved BILAG score because there was essentially no change from that extremely active disease.

DR. ISENBERG: In practice, I don't think that really works.

DR. FELSON: But theoretically, that could work, right?

DR. ISENBERG: It could work. What it does, though, what the BILAG system does is it enables you to see, if you like, at a glance what is going on in particular organ systems at a particular time.

DR. FELSON: But what you are describing is a state, but the way you put it up in terms of how things are

measured is transition. So, the other issue with that is if you don't see the patient for six months or you see the patient one time and another doc sees the patient the next time, they don't have any sense of what the transition accurately ought to be.

DR. ISENBERG: Well, as I have said, the BILAG score refers to what has happened in the previous month. An assumption is made, if you like, that these patients are going to be followed up at times that are appropriate to their degree of sickness.

DR. FELSON: If it is a transition question, then, how do you compare that previous month to what time in the past is this compared to, is that defined?

DR. ISENBERG: Well, it is simply compared to what was going on at that time the patient was previously seen.

DR. FELSON: Let me also ask. You showed its lovely validation and reliability stuff you guys have done, and it has been impressive to read about it also, but is there data here on the relative sensitivity to change of any of these instruments or on their redundancy?

DR. ISENBERG: That has been done and it has been published as a comparison. All three were shown to be very sensitive to change. When I say "three," I mean the BILAG, the SLAM, and the SLEDAI were compared in a study which we published four or five years ago, and I have a copy of that

with me if you want to see it.

DR. FELSON: The correlation matrix you showed suggested there is a lot of redundancy, but one of the questions you brought up very nicely at the beginning was how I think from a validation perspective, one would be concerned about the concept of content validity meaning that there is a broad spectrum of disease activity, and if one wishes to sample from all of those different elements of the spectrum, and one of the ways you did that very nicely was to comment on the use of SF36 as a measure of sort of the other dimension of disease activity.

One of the questions that comes into play with the instruments that you commented on is whether some of them sampled from certain domains of disease activity that others don't sample from, and whether those domains of disease activity are important with respect to measuring activity and with respect to change in activity.

Therefore, one of those instruments might be preferable to another. Do you know if one can evaluate that by doing factor analyses, one can evaluate that by just looking at elements of the instruments to see which are not incorporated in some versus another?

Do you have a sense from the instrumentation of what elements are involved with one and not the other?

DR. ISENBERG: I think each of these different

instruments have tried very hard to provide the interested lupologists with a comprehensive overview of what is going on. Now, whether you divide lupus activity up into eight systems, as BILAG does, or nine as SLEDAI does, I think it is just a question of cutting up the cake in a slightly different way.

I really don't think any of them are missing anything terribly major, if you like. I think they are looked at from slightly different perspectives. What I think has been very interesting for us to look at over the years is the realization that although we have approached this problem from very different philosophical standpoints, these indices seem to come up with pretty similar numbers or pretty similar ideas of activity whichever way we looked at it.

DR. FELSON: The other question is you commented on the difference between SF36 and SF20 and noted prominently the importance of fatigue questions. What does SF36 contribute to lupus evaluation in terms of either sensitivity to change or information over and above that single question on fatigue?

DR. ISENBERG: Well, I think those questions are actually being asked right now. We, for example, are doing a five-year prospective study in a sense to try to answer that precise point. Indeed, what I showed you was a cross-

sectional analysis of the data over the first year that we have on these 140-odd patients that we have entered into this cohort.

We are now into the fourth year of the study, and I think we will be able to answer that question definitively for you in about a year's time.

DR. ABRAMSON: I think we will have to move on now. Thank you, David.

Next is Dr. Petri from Johns Hopkins, Definition of Flare and Responder Index.

Definition of Flare: Responder Index

DR. PETRI: Ten years ago several of my colleagues and I got together to define flare in lupus.

[Slide.]

We defined it as a change of 1.0 or greater on a zero to 3 visual analog scale of disease activity. We then looked at the first 185 patients who had entered the Hopkins lupus cohort and found that 53 percent had this pattern of flare.

The incidence of flare was 0.65 flares per patient year of follow-up, and the median time from the first study visit to a flare was 12 months. We thought this was very important information in the design of clinical trials that wanted to look at a decrease in flares as an outcome.

[Slide.]

More recently, with my colleague Susan Barr and Abraham Zonana-Nocach, we have determined that rheumatologists can look at these disease activity measures over time in individual patients and determine several patterns of disease activity. This can be done with excellent agreement using the visual analog scale of the physician's global assessment or using the SLEDAI as described by David Isenberg.

[Slide.]

I want to show you one of these patterns that we think is very important. We call this the relapsing remitting pattern basically, borrowing from the MS terminology, but it is fine to call it a flare, as well, and you can see in this patient using either the physician's global assessment or the SLEDAI modified to remove complement and anti-DNA descriptors it is very easy to see this pattern of relapse-remission, relapse-remission going on for many years in a patient who is under treatment.

This kind of patient poses some problems for those of us who are thinking about enrolling patients in clinical trials. If you enroll this patient in a clinical trial at this time, the relapse time, obviously, the natural history is for that patient to get better and for there to be many months perhaps before the next relapse.

On the other hands, you could enroll a patient in

a clinical trial at this time of remission and have a problem. There may be many months before that patient's natural history will allow her to have another flare.

[Slide.]

Here is a second pattern that I think people weren't quite as aware. This is the chronic active pattern. This is a lupus patient who always has disease activity regardless of whether we use a visual analog scale or the modified SLEDAI. This patient never achieves "no activity."

I think this, in fact, is a patient who might be better entered into clinical trials because it is a little easy to show improvement in someone who has chronic activity.

[Slide.]

There are some patients who have a period of long quiescence although we prefer not to use term "remission," because as you can see in this patient, after three years of "no activity," there is again evidence using either the visual analog scale or the modified SLEDAI of SLE activity.

[Slide.]

How often do patients show these patterns? In our most recent analysis, we almost completely duplicate our results from 10 years ago. Fifty-one percent of patients show a relapsing-remitting or flare pattern, 80 percent at some point have a period of chronic activity, and only 17

percent have a period of long quiescence.

This does not add up to 100 percent because patients can change patterns. I think that is going to be a very important thing to look at in clinical trials especially of stronger agents, such as intravenous cyclophosphamide, do the stronger agents change the pattern from flare and chronic activity to one of long quiescence.

[Slide.]

When we look at the second index, the modified SLEDAI, we don't see exactly the same percent of time in these different patterns, and I think we all recognize this, that some instruments capture disease activity better in some organ systems than in others.

[Slide.]

We were very interested in predictors of the flare pattern, the relapsing-remitting pattern, and we found that there are some demographic issues that are of importance. Female patients with lupus are more likely to have a flare than males. We also found a hormonal issue in women in that postmenopausal women were less likely to flare than premenopausal women. These kinds of things need to be taken into account when patients are enrolled in clinical trials.

[Slide.]

We, as everyone else in this field, would like to find surrogate markers, but we have not found that

serologies are helpful. Specifically, when we look at C3, C4, and anti-double-stranded DNA by crithidia, we cannot find that changes are predictive of a flare in the next three months.

There is great controversy in this field. Some people believe that a Farr assay might be a better predictor of flare than crithidia. Some people believe that these assays, in order to be predictive, need to be done on a monthly basis. We do the assays on a quarterly basis.

[Slide.]

I want to show you, though, how complicated it is to look for surrogate markers. We have looked very carefully in our longitudinal database at whether serologies correlate with disease activity at the same visit. In fact, they do, but in such a weak way that clinically it is not useful.

For example, if we look at C3, the very lowest values of C3 do associate with higher levels on the visual analog scale or on the modified SLEDAI. When you consider that the physician's global assessment goes from a zero to 3 scale, and 1 is mild, you can see that there is very little spread here. Similarly, on the SLEDAI, very little spread.

[Slide.]

For C4, we appear to have found a U-shaped relationship where at the very lowest levels of C4 are

associated with higher levels on the physician's global assessment, but so are the very highest levels of C4. Looking at the SLEDAI again, that same U-shaped relationship, the lowest levels and the highest levels are associated with greater disease activity.

[Slide.]

Again, for anti-double-stranded DNA, the very highest levels of anti-double-stranded DNA are associated with higher levels on the visual analog scale or on the modified SLEDAI, but the spread is very small, so this is not clinically useful.

[Slide.]

We have actually constructed random effects models to look at this, and all these different serologies are significantly associated with disease activity at the same visit. The problem has been when we interpret these models, they would not be clinically useful and probably would not be useful in the research arena either.

I want to go over with you in detail the results for C3. The average value for the beta in the population is minus 0.0029. What does this mean? It corresponds to a change in the expected physician's global assessment of 0.029 for each 10-point change in C3, which would be a large change. Remember, the physician's global assessment is on a zero to 3 scale, so this amount of change is infinitesimal.

The problem, though, is that people vary in the lupus population with respect to beta, so considering the middle 95 percent of the values of beta in the lupus population, some are as low as minus 0.01, but some are actually in the opposite direction and are as high as 0.0095. The clinicians in the room understand this. This is why in some patients you keep flow sheets. In some patients, these are useful, in other patients they are not.

[Slide.]

I wanted to show you that anti-double-stranded DNA is not a good surrogate marker even if we limit its use to those patients who have shown the capability of making anti-double-stranded DNA. Changes in anti-double-stranded DNA even in this subgroup do not associate with the probability of flare or with a change in the physician's global assessment.

[Slide.]

As part of the SELENA trial, which is the safety of estrogen in lupus national assessment trial, Dr. Buyon is the co-principal investigator, and we have multiple enrollment sites, many of whom are in the audience.

[Slide.]

We took a very tough committee look at both flare and disease activity. We decided to use the SLEDAI as the disease activity measure in this trial using the same

descriptors and the same weights, but we made modifications to the definitions of the descriptors to ensure that we truly captured disease activity.

[Slide.]

In the SELENA trial, we defined flare. This was defined by our investigator group in meetings, and we defined both mild, moderate, and severe flares by consensus.

[Slide.]

I want to show you the mild, moderate flare definition. It included a change in SLEDAI of 3 points or more, new involvement or worsening of these disease activity descriptors, the physician increasing prednisone, the physician adding an NSAID or Plaquenil for disease activity, or a change in the visual analog scale because occasionally, even with our modifications in SLEDAI, we have a flare in an organ system that SLEDAI would not measure.

[Slide.]

We also defined severe flares, a change in SLEDAI to greater than 12, new or worsening involvement of these descriptors, an increase in prednisone to greater than 0.5 mg/kg/day, addition of an immunosuppressive drug for active lupus or hospitalization for active lupus, and an increase in the visual analog scale to greater than 2.5 on a 3-point scale.

[Slide.]

We then basically did a reliability validity study. It had two purposes - to determine the reliability of the SELENA flare definition and to determine the reliability of the SELENA SLEDAI. We actually did two reliability studies with physician training in between.

[Slide.]

The first study involved 7 paper patients. These patients were derived from our Hopkins lupus cohort, and six of the participating sites looked at these paper patients.

The inter-class correlation coefficient for the SELENA SLEDAI was superb. For severe flares, the kappa for agreement was also substantial. Our problem was that for mild-moderate flare, the amount of agreement was unacceptable.

[Slide.]

Physician training occurred by our comparing our answers, and we then did a second reliability study. This time the patients actually came from the SELENA trial, the same six physician raters. Once again, the SLEDAI had excellent inter-class correlation coefficient, and for severe flare, the kappa remained substantial. We were able to show that for mild-moderate flare, with physician training, the kappa improved dramatically.

Remember that the people enrolling in this trial are all lupus experts, but it still took training for us to

agree on how we define mild-moderate flares.

[Slide.]

I think the important message for industry and for everyone on the panel is that physician agreement on the SELENA SLEDAI is excellent without any training, but as we get into this definition of flare, we all seemed to agree on what a severe flare is, but we need to be trained if we are going to capture mild-moderate flares reliably.

[Slide.]

Now, there are other issues that have come out in the SELENA trial that I want to present to you this morning because they are going to be important in other trials, as well. I want to show you the problem in comparing mean disease activity scores versus looking at changes in flare rates. These analyses were done by Mimi Kim, who is the biostatistician for the SELENA trial, with the help of Mary Lou Skovron, who is a member of this steering committee.

[Slide.]

When we look at patients who have had a severe flare, a mild-moderate flare, or no flare using the SELENA definition, and we then look at what their SELENA SLEDAI scores were, in both the estrogen replacement therapy trial and the oral contraceptive trial, you can see that the SELENA SLEDAI gives higher scores in patients who had a severe flare versus a mild-moderate flare versus no flares

as you would expect.

Now, obviously, severe flares are rare, mild-moderate flares are more common, but there are, of course, the patients who have that chronic disease activity, as well, who are not flaring.

[Slide.]

So, if we look at the end of the trial, these are models, SELENA is obviously ongoing, and we look at the expected difference in the mean SLEDAI between treatment arms, we are going to have problems.

Let's assume, for example, that there is a 9 percent difference in severe flare rates and a 20 percent difference in the mild-moderate flare rates. Would looking at just SELENA SLEDAI capture this difference?

The expected difference in the SELENA SLEDAI in the ERT trial, if these percentages remained true, would only be 0.62 between the two groups; in the OCP trial, only 0.46 between the two groups. These are very small, expected differences in SLEDAI and what is in fact a very large trial. Both of these trials are very large.

So, this is somewhat daunting. I don't think we can expect one instrument, such as the SLEDAI, to be able to capture differences in disease activity as well as differences in flare, and I am going to suggest to you that we have to be very careful to include flare as an outcome,

as well.

[Slide.]

We think that it is very important that this committee consider responder indices. With the help of IDEC, a group of us, including Dr. Buyon, Ginzler, Kalunian, Merrill, and Wofsy, have started work on a responder index that we call the RIFLE. This is a work in progress, and we welcome the help of everyone.

[Slide.]

This is a little bit hard to see, but what I want to go over with you is the philosophy of the RIFLE. There are many different ways to decide at the end of the trial that a patient has either responded or not.

We can decide that on an overall basis, and we have decided to define that a patient has responded if organ systems stayed the same or got better, and no organ system got worse, but there is another way to determine that someone has responded during a trial, and that is by organ system.

[Slide.]

So, for example, you could at the end of a trial actually come up with a count of how many people with renal disease responded, how many people with cutaneous disease responded, and this might be very important for some drugs, a drug such as Plaquenil, for example, which is expected to

help cutaneous and musculoskeletal activity, but would not be expected to help CNS or renal activity.

There is a third way to look at response in a clinical trial that was suggested by David Wofsy. Basically, physicians know which organ system is most important in that patient, and is the reason that they have entered the patient in a clinical trial.

So, for example, if a patient is enrolled in a clinical trial because she has renal disease, and her skin and joint manifestations improve, but her renal disease did not, that patient is not a responder in the organ system that the physician ranked as most important.

So, there are at least three different ways, and perhaps others, in how to define a patient or an organ system or a rank-order has responded in a clinical trial.

[Slide.]

This is an example of how the RIFLE is organized. We think it is very important that there be definitions for everything. We learned this when we did SELENA. I think many of the industries that have been working in lupus trials have learned this, as well.

Basically, for each descriptor in an organ system, we think it is important to actually define what is worsening, what is no change, what is a partial response, what is resolution. These are still being worked on by the

committee and by members of the SLICC group, but hopefully, we can reach a consensus on how to define these.

[Slide.]

Now, David Isenberg mentioned that the patient cares most about fatigue, and I think everyone who treats lupus patients will agree. One problem we are having in clinical trials right now is that fatigue is of two types in lupus patients. There is an acute component of fatigue that seems to occur during flares, and there is a chronic component of fatigue that seems in many centers to associate very highly with fibromyalgia.

[Slide.]

I want to show you in our longitudinal study that a whopping 29 percent of our lupus patients meet ACR criteria for fibromyalgia on the basis of number of tender points, chronic fatigue, and pain.

[Slide.]

I want to show you the effect this has in a clinical trial setting. I want to thank Genelabs for allowing me to show this. This is in the qualifying visit of the Genelab study showing you the SLEDAI disease activity measure on this axis and the fatigue severity scale on this.

You can see how most lupus patients cluster at the worst end on the fatigue severity scale, but most importantly, there is no association of disease activity

with fatigue when patients enter this trial. So, this is a major issue, if improvement in fatigue is an outcome measure it may not actually be an outcome measure closely related to disease activity.

[Slide.]

It does not make any difference using the Genelabs data whether SLAM is used. At the qualifying visit, SLAM does not associate with fatigue either.

[Slide.]

Now, I know there are many industry representatives in the audience today, and I wanted to end with a little bit of philosophy about why you should care about lupus clinical trials and why you should get involved.

We are ready for you. We have the disease activity measures. Some of them require refinement, but they are ready. In other words, we built the ballpark, now we want you to come, and the reason you should care is that lupus still has a major problem with mortality in the United States.

This is the study recently published in Arthritis and Rheumatism from the Mayo Clinic, Sherri Gabriel's group, showing at 10 years that there is less than 80 percent survival, and this is in middle-class caucasians. You can imagine what survival data are like in academic centers in inner cities.

[Slide.]

But here is the second reason why you should get involved. Dr. Gabriel's study showed a 3-fold increase in the incidence of lupus since what I will call the early era to the current era. This is using exactly the same criteria for the classification of lupus, so this is not due to diagnosis of milder cases. In fact, the frequency of renal disease is exactly the same in the current era as it was in the fifties to seventies.

So, this is a major problem in our population where now 1 out of 800 middle-class caucasians in Rochester having lupus, and if we had these kinds of excellent epidemiologic data to reflect the other racial groups in the United States, I think we would find it was even more of a problem.

I thank the committee for their attention.

DR. ABRAMSON: Thank you, Dr. Petri.

We have time for maybe one question.

DR. FERNANDEZ-MADRID: Your data on anti-DNA antibodies is mainly on crithidia?

DR. PETRI: That is correct, and that is a limitation. It is possible that the Farr assay might be more responsive to disease activity changes.

DR. FERNANDEZ-MADRID: Yes, I think we have to have an open mind on that, and also the Farr assay has

problems depending on the antigen involved. So, I think further information is needed in this area, it seems to me.

DR. PETRI: I think everyone agrees.

DR. ABRAMSON: Thank you, Dr. Petri.

The next speaker is Dr. Dafna Gladman from the University of Toronto to discuss damage and drug toxicity.

Damage: Drug Toxicity

DR. GLADMAN: Well, I have a task that is a little bit easier because the concept of damage in lupus is a little bit less controversial partly because the groups involved in the development of the concept of damage and the damage index got together early enough in the course without having individual indices that were already attached.

[Slide.]

So, you ask yourself why do we need a damage index in lupus. Despite the fact that there is still a significant mortality, in fact, patients with lupus have a 3 time risk of mortality than the general population even in the 1990s, there has been a reduction in mortality, and the survival rates that were quoted by Marilyn Schulman in 1955 of 50 percent in five years, are now close to 70 percent in 20 years, therefore, mortality is no longer an appropriate outcome measure in lupus.

Moreover, disease activity may result in organ damage, and we are all aware of the fact that when a patient

has lupus nephritis, they on occasion go on to dialysis and renal transplant. When a patient has neurological involvement, they may end up with a seizure disorder or the result of a stroke, again resulting in an organ damage.

The management of patients with lupus would then include not only the prevention of death, but also the reduction of morbidity which results from those organ damage items that we are going to discuss.

So, we felt that a method to estimate morbidity, i.e., damage, was necessary.

[Slide.]

Now the first group to get together was the Conference on Prognosis Studies in Lupus, which took place in Toronto in October of 1985, and again you will recognize some of the individuals like Matt Liang and Peter Schurr, and others, including the late John Decker, who were included in the group that actually developed the SLEDAI, the SLE Disease Activity Index.

During the same conference, the participants were asked to include in their thinking the concept of patient perception that David Isenberg has already talked to us about, and the concept of damage.

[Slide.]

During the 1985 conference, three steps were taken to try and arrive at a damage index. The first step was to

review a list of items that were considered to be damage resulting from the disease or its consequences.

The participants then ranked the damage items in a similar way that was done for the activity items. Unfortunately, for the damage, we were not able to find a point between items that should be included and items that could be excluded from an index, and therefore, the variables were defined just based on the ranking system.

[Slide.]

During that 1985 conference, the definition of damage that evolved was that it was an accumulated end organ irreversible effect of persistent disease activity, drug therapy, or intercurrent illness.

[Slide.]

Now, in 1990 or 1991, the Systemic Lupus International Collaborative Clinic group was formed including individuals from Canada, Italy, the United States, Great Britain, Sweden, Mexico, and Australia, and this is the SLICC group, and members of the SLICC group went on to develop a damage index which could be tested for validity, reproducibility, and sensitivity to change, and the generation of the index was by a nominal group process based on the 1985 conference template.

[Slide.]

During that meeting in Boston, 20 patient profiles

containing damage concepts were assessed. An analysis revealed patient variability which was what was desired, but it also revealed significant variation among raters, and therefore a review of the items and definitions had to be carried out again.

[Slide.]

In order to generate the index, we again went through the process of assessing the items to be included, define each item for ascertainment, agree on the item and the definition, and eventually came up with an index of 12 organ systems, and content validity was demonstrated at least within the SLICC group.

[Slide.]

The damage index was defined based on an irreversible change since the onset of lupus, ascertained by clinical assessment, so that it would be useful to anybody who wished to use it as opposed to only areas where very high technology could be used, and the other important item is that for most items, unless stated otherwise, they had to be present for at least six months.

[Slide.]

The end result of that initial SLICC meeting was that here is the damage index with the 12 organ system and the items that were included in each area.

[Slide.]

I will show you examples of a couple of them. The ocular, for example, if a person had a cataract ever, so here there is no need to be a six-month window, a cataract ever gets a score of 1. If there is retinal change or optic atrophy which is present for six months, gets a 1, so therefore, the ocular system gets a total score or a possible total score of 2.

[Slide.]

Looking at the neuropsychiatric system, a persistent cognitive impairment or major psychosis, again six-month window, seizures requiring treatment for greater than three months gets a 1. A CVA, the first one gets a score of 1. If another CVA occurred six months later, it now becomes a 2, and so on. So, the CNS system gets a total possible score of 6.

[Slide.]

In the renal system, an estimated or measured GFR less than 50 percent scores a patient 1, and, of course, that has to be there for six months, so as not to just reflect disease activity.

The presence of persistent proteinuria of 3.5 grams or more again scores a 1. However, if a patient has gone into end-stage renal disease, regardless of dialysis or transplantation, they are scored as 3, and they don't get scored for those.

[Slide.]

Now, in order to validate the index, we asked the members of the SLICC group to provide four patients who have had disease duration for at least five years, and of the four scenarios based on their own patients, two had to have active disease and two had to have inactive disease at two time points, one with and one without damage.

[Slide.]

So, 19 physicians completed damage index on 42 case scenarios based on this accumulated information from the various centers. We looked at time, one or two, the damage and activity, and found that they were all statistically significant and that interaction of time with activity and damage with activity was significant, however, the effects due to physicians were small, in fact, it was 1 1/2 percent, suggesting that the physicians scored those scenarios in a very similar way.

[Slide.]

Here is what happened with the physicians' score over time. In patients that were active, but stable, there was very little change in damage. Patients that were inactive and stable, very little change in damage. However, patients that were thought to have an increased damage, whether they were active or inactive, clearly reflected that change based on the scores of the scenarios for the 19

physicians.

[Slide.]

In addition, we asked our local rheumatologists at each of our centers of the SLICC group to review the index, and there was major agreement with the index by the local rheumatologists. There were minor changes that were incorporated in the final index, and we were now ready to test for validity.

[Slide.]

One of the things we wanted to do was to confirm the reproducibility of the SLICC in live patients as opposed to just the paper exercise. So, we got together in Toronto to test the reliability and validity in the assessment of live patients and to correlate the presence of damage score with disease activity scores.

[Slide.]

So, 10 patients representing a spectrum of damage and activity were recruited from the then lupus clinic in Toronto. Ten physicians from five countries representing 10 lupus clinics assessed the patients, and remember, these 10 physicians from five countries represent different concepts and also different health care systems.

The order of the physicians and the patients was randomized according to a Youden square design.

[Slide.]

Each patient was examined by six different physicians, and each physician examined six different patients, and the SLICC ACR damage index was used to assess damage while the SLEDAI was used to assess the activity.

[Slide.]

These are the average scores for the measures.

[Slide.]

The reliability of the SLICC score is shown here where the patients contributed the variance, and this of course was by design, but you can see that the order of the assessment or the physician had no effect on the variability here.

[Slide.]

Likewise, for the SLEDAI, the patients were selected appropriately, and there was no patient order effect.

[Slide.]

So, we have demonstrated that the SLICC ACR damage index detected differences among patients, and that the SLEDAI also detected differences among patients, and that there was no detectable observer difference with either the SLEDAI or the SLICC ACR damage index, and there was no order effect.

[Slide.]

The conclusion from it is that physicians from

different centers can use the SLICC ACR damage index to assess damage in patients with lupus with a spectrum of both disease activity and damage.

You will notice that I changed the title. It is not just the SLICC damage index, the damage index was then approved by the American College of Rheumatology and adopted as the damage index for lupus, and in fact is published.

[Slide.]

We further went on to see whether the damage index could actually discriminate between patient populations, and we asked physicians from our group again to provide us with information about patients that they are following in their practices, and we had a database of 1,297 patients.

You can see that there is representation from Baltimore, Birmingham, England, Halifax, Lund, Sweden, Middlesex, England, Montreal, Canada, New York, and Toronto showing that the majority of the patients were females, and the majority of the patients were caucasians, although you can see that in the American centers, there was a higher representation of non-caucasians.

The mean age at presentation was similar, although again there was a slight nonsignificant variation.

[Slide.]

Now, you can see the ranges of the damage index scores between the centers. The initial score in Baltimore

was quite low, on the other hands, Birmingham had an average score of 1.24, New York had an average score of 2, and Toronto was similar to Birmingham. The score in Lund, Sweden, was very similar to the one in Baltimore. This is the initial score.

[Slide.]

Five of the centers were able to provide us information that would allow us to look at the change in the SLICC damage index over time, and this would be within the first two years of disease, this would be after five years, and this would be after 10 years, and you can see that regardless of the center, there is an accumulated damage which occurs over time, suggesting that the SLICC does record changes over time.

[Slide.]

Now, is it an important concept to recognize? The answer is yes. If you look at patients who ended up dying in the course of these studies, the score for the damage index for the individuals that remained alive was lower than the score of individuals who ended up dying. This is the initial score, and this was highly significant, suggesting that a damage index is in fact predictive of mortality.

[Slide.]

So, although there are some demographic variations, we noticed that patients in Baltimore were

younger than those in other centers.

[Slide.]

The SLICC ACR damage index did measure damage, the damage reflected accumulated change over time and higher damage scores were documented early in their course in patients who went on to die.

[Slide.]

Now, there have been further validations of the SLICC ACR damage index. David has already shown some of the information relating the damage index through patient perception, but the other study, they did show that renal damage at the year was predictive of end-stage renal disease, pulmonary damage was predictive of death, and that Afro-Caribbean and Asian patients accumulated more change, more damage than caucasian patients.

[Slide.]

The LUMINA study used the SLICC ACR damage index and showed that from the outset of their disease, diseased patients experienced more active disease and more accumulated damage than survivors, very similar to our international study.

[Slide.]

The Montreal cohort, which is Paul Fortin's cohort, showed that the SLICC ACR damage index scores predicted poor outcome in patients with lupus both in terms

of death and in terms of hospitalization.

[Slide.]

The Dutch treatment study showed that the SLICC ACR damage index has the capacity to detect change over time, again confirming our international study. Further, it was found to reflect the impact of cumulative disease activity particularly in renal and hematologic, and cumulative doses of prednisone, and they felt that it was a useful outcome measure.

[Slide.]

So, what I hope I have left you with is the conclusion that the SLICC ACR damage index is a valid measure of damage, and that it is a useful measure in studies, and its use in clinical trials may be several fold.

First of all, it can be used as a descriptor of the patient in addition to these activity measures. Patients perhaps should be stratified by the presence of damage.

[Slide.]

In fact, one could use the damage index as an eligibility criteria for clinical trial, because if a patient has already had a major degree of damage in either kidney or brain, they may not be appropriate for clinical trial in a medication which is supposed to prevent damage, and finally, the damage may be used as an outcome measure in

a clinical trial, not perhaps for a medication which is supposed to turn off inflammation, but as a long-term outcome in the sense that if the drug is good in terms of controlling disease activity, then, it should demonstrate the ability to prevent the accumulation of damage either globally or within a particular organ system.

[Slide.]

Now, I was also asked to talk about toxicity, and I actually decided not to bring a lot of slides on toxicity because both the OMERACT group and the NIH have already developed a toxicity index which is a 20- to 30-page document, which reflects toxicity from any type of drug.

What I wanted to mention, and I apologize for the quality of this particular slide, I guess it can't be focused any better, but what this slide represents is a table of toxicity from the drugs that are already available and used in patients with lupus, so there is a toxicity that arises from nonsteroidal antiinflammatory drugs that would require assessment prior to using the drug, monitoring using the drug both in terms of clinical assessment and laboratory assessment.

This is basically taken from the guidelines for the monitoring of lupus and the management of lupus which the American College of Rheumatology has recently adopted and hopefully will be published soon, but again reflecting

toxicity from nonsteroidal antiinflammatory medications, antimalarial drugs, and immunosuppressive cytotoxic drugs.

Some of these toxicities are already included in the SLICC ACR damage index.

[Slide.]

If we go back, you can see that some of the ocular toxicity, for example, cataracts, would reflect changes that have occurred from steroid therapy, and some of the retinal changes may in fact reflect changes that have occurred from the antimalarial therapy.

Likewise, some of the changes, for example, in the musculoskeletal system, such as osteoporosis and avascular necrosis, would account for some of the drug toxicities, such as steroid therapy, premature gonadal failure might account for drug toxicity such as cyclophosphamide, diabetes may account for drug toxicity such as steroids, and malignancy may account for some of the toxicity from some of the drugs we are using.

Since the SLICC ACR damage index does not attribute, it only records damage, whether it is the result of the disease process itself, whether it is the result of drug toxicity, or whether it is the result from intercurrent illness, since the onset of SLE, it would accommodate some of the toxicity that we would incorporate in toxicity record, but as I say, since there is already an acceptable

toxicity record, I did not think that it was appropriate for me to discuss this in more detail, but I certainly would be pleased to answer questions.

DR. ABRAMSON: Thank you very much. I just had one question. What is the sensitivity to the SLICC in determining damage in the clinical trial over a shorter period of time, six months and twelve months? Does it have any utility versus the disease-activity indices that we have heard about?

DR. GLADMAN: Remember that the damage index reflects nonreversibility. So, you certainly would not expect a sensitivity in the same way that you expect a sensitivity to change in an activity index.

By definition, you cannot use it in anything less than six months because in order for an order to be scored, it has to be present for these six months which means that it probably doesn't make sense to use it in anything that is less than a year, because if your trial is six months and you expect the change to occur in six months, then, we are talking about an instrument that would be useful for anything over a year period.

Now, you do not expect the damage index to go down, so what you would be expecting from a damage index is not to go up. So, if you are expecting from a drug to be sensitive or the damage to be sensitive vis-a-vis a drug

trial, the sensitivity should be that there would be no increase in the SLICC ACR damage index over the year.

So, you wouldn't use that if you are trying to justify changes in disease activity, but you would use it as an outcome of the clinical trial or an outcome of the drug to record lack of increase. That is the way we would interpret it.

DR. ABRAMSON: Dr. White.

DR. WHITE: Dafna, I have two questions. The first is a follow-up on that.

If you wanted to use this as a descriptor of the severity of the patients or stratifier of the severity of patients, you showed us a really striking correlation between years after first appearing in a clinic and this score.

Does this score offer anything more in terms of initial characterization of severity of disease damage than just years after follow-up?

DR. GLADMAN: I am not sure that I understand the question.

DR. WHITE: The question is it would be important to be able to stratify and identify damage in patient populations when they are entered in clinical trial, and that is what you have tried to do, but you also showed us that there is a very good correlation between what this

score is in a variety of centers and the number of years after a patient showed up for care.

So, I wondered, is just duration of disease as good an initial characteristic of disease damage as this index.

DR. GLADMAN: Okay. So, first of all, if you just look at the mean of the damage index within the first two years, you can see that there is a variability, right? The mean score for Baltimore was 0.36, the mean score for New York was 2.07, and if you remember, the distribution of patients in those two centers was very similar, almost 50-50, so it shows you that if you wanted to characterize the patients within the first two years of disease, I mean you could characterize it on onset, which would be zero, or at six months, but it shows you that you can say that yes, the group in Baltimore was less damaged within their first two years of disease than the group in New York.

So, if you are including patients in a study, you can say that you want their damage index to be less than or not exceed whatever, or you want to have patients where you want to use the mean SLICC as comparison between two populations of patients.

DR. WHITE: Were those numbers that you are showing us based on the disease onset defined in some way or the time they showed up in that clinic?

DR. GLADMAN: These are patients who showed up in the clinic and had their SLICC scores done for the period two years.

DR. WHITE: It may be that patients show up in Baltimore sooner than they show up in New York. That was all. It is just a question about --

DR. GLADMAN: No, this is from the time of diagnosis. These people were diagnosed in the centers.

DR. GINZLER: Some of the patients weren't seen at the time of diagnosis. It was from the time of entry.

DR. GLADMAN: So, that may reflect the difference between Baltimore and New York.

DR. WHITE: It is just a point, what is easiest to do, and if this is really by time of entry rather than time of disease onset, it may be useful.

Has anyone looked at duration from time of disease onset, not diagnosis, but what seems to be onset, and their characteristics. Maybe that is an easier way to do it. That is all I am asking.

DR. GLADMAN: This study is actually being done at the moment, at the University of Toronto, but the information that you have here.

[Slide.]

For example, for Lund, these are 12 patients that were entered at the time of diagnosis. This is in their

first two years of disease. This is at five years of disease, and this is 10 years of disease. You can see that at least for this population, there is a clear difference, and it is possible that the Baltimore group -- Michelle?

DR. ABRAMSON: Excuse me, Dr. Gladman. Excuse me one second. I think what we will do is we will have to come back to this during the panel discussion, if you just want to finish your thought, and then we will go on to the next speaker.

DR. GLADMAN: Well, if Michelle entered patients in the same way that the Lund group entered them, then, you can see that the first two years, that these two groups are very similar.

DR. ABRAMSON: Thank you very much.

DR. GLADMAN: But I think it can be used to compare patient populations.

DR. ABRAMSON: Our next speaker will be Dr. Vibeke Strand from Stanford University on the OMERACT recommendations.

OMERACT

DR. STRAND: Thank you.

[Slide.]

I wanted to say that I appreciate very much being invited to participate in these discussions and that they do follow from I hope the efforts with OMERACT, which have been

efforts to be inclusive and to build upon the extensive amount of work that you have heard presented here already this morning.

I wanted also to point out the Journal of Rheumatology, whether it was wittingly or unwittingly, very nicely published the proceedings of the OMERACT lupus module in this month's Journal of Rheumatology starting on pages 490 and further, and what I wanted to do today was to sort of review for you a summary of what occurred at that meeting.

[Slide.]

We all know that analyses of outcome in lupus are complex. I think that Dr. Schwieterman was very correct this morning to say that it was really the disease that was a problem, and not the people working in the disease.

We have had a lot of difficulty in trying to define what our goals are in terms of randomized clinical trials in lupus because there have been so few. We have obviously the example of the nephritis studies, but, in fact, people are now trying very hard to look at treating lupus in the context of not just nephritis, but the other organ system manifestations, or, in fact, in looking at organ system manifestations to the exclusion of nephritis.

Of course, right now much has been learned from the cohort studies and a variety of instruments that have

been developed in that context, and we are now seeking to apply these to randomized clinical trials.

There are quite a few biologic agents that either have been or are starting to go into trials in lupus. I think it is a very exciting time, as Dr. Petri just pointed out, we have now built the field.

[Slide.]

There are a few pharmaceutical agents, both new and old, that are being applied in different ways and looked at in lupus, and again, much of this data will be of great interest to us.

[Slide.]

We have discussed a variety of the disease activity indices. In fact, there are five of them. There is also the LAI, which is not listed here, but these five have all been utilized in clinical trials although the data have not necessarily been published, and unfortunately, because of that, we don't have the information as to their reliability and sensitivity to change in a randomized clinical trial.

We look forward to publication of a couple of recently described clinical trials because I think this will help us considerably in this field.

[Slide.]

This is just a comparison of these different

disease activity indices. I think it is important to note that each of them does, in fact, offer some benefits. They are all different in the sense that they were developed in different ways, and increasingly now they are scored with the idea that each one is looked at for disease activity in the previous month.

[Slide.]

What is useful is that there is a computer generation of all five of indices which will allow you to use whichever one you would like to use, and then have all of the others applied.

I think this has been of great benefit in terms of the clinical research efforts of the SLICC group that you have heard from this morning, as well as ESCICIT and Euro-Lupus groups, and it would allow us to exchange information between our different groups in terms of our different clinical trials and in terms of our clinical products.

From that point of view, Stefano Bombardiere and his group has made this computer program available free of charge to anyone who wishes to use it, and is encouraging people to use it within the academic community, and it is offered also to those of you who are doing or sponsoring randomized clinical trials for a very reasonable donation to continue the efforts of ESCICIT and Euro-Lupus.

[Slide.]

We have looked about flares in this past discussion, and there are several different definitions of flare. Fortin's group has a specific definition in terms of major and minor flares, whether hospitalizations or emergency room visits are required.

Ginzler and group have published using the SELENA SLEDAI definition of flare, which is not dissimilar to what Dr. Petri was just describing, Dr. Gordon's group using the BILAG in terms of the A or the change to a B from D or E, as Dr. Isenberg presented earlier this morning, and in a small study of patients, Fitzgerald using the SELENA flare definition with Ken Kalunian's group.

What is important here to me is that these are rather similar incidences of flares be they major or be they minor, in other words, however we go about defining this, we are coming up with a remarkable consistency across these different patient cohorts for the incidence of flare.

[Slide.]

The damage index was just discussed in great detail, and I think it is a very important assessment that must be included in clinical trials. It obviously may be most useful in terms of stratifying patients at study entry, and it may also be a very good way to define a definition of treatment failure.

[Slide.]

Health status and health-related quality of life has been looked at in lupus increasingly using the SF36 although published data with the health assessment questionnaire and the SF20 have shown them also to be very useful in lupus populations.

In general, because of the component of fatigue that seems to be quite present in lupus whether it is due to the disease itself or associated fibromyalgia, it has been decided that the SF36 is a preferable instrument.

What we know about the SF36 as a generic instrument is that it really is very helpful for us as rheumatologists in whatever disease we are studying to allow comparison of our rheumatologic patients to other patient populations.

The SF36 is undergoing extensive cross-cultural translations and validations, and there is a large patient normative database at this present time, and it has been shown in a variety of cohort studies that the SF36 is sensitive to change over time in lupus.

[Slide.]

What we also know is that baseline subscale scores in lupus patients tend to be very low. This has been published in cohort groups from Canada, Norway, and the UK, United States, and that the subscale scores tend to look like women of age and sex matched, who have serious medical

problems, such as diabetes or coronary artery disease.

We also know that lupus affects psychological social well-being, as well as physical well-being.

[Slide.]

A variety of groups have shown that, in fact, changes in the SF36 are reflected also by changes in disease activity measures, that, in general, decreases in disease activity show improvement in physical functioning, bodily pain, and general health profile of the SF36.

This was published by Gordon and group, but also, Fortin and Ginzler, and so on, have shown that increasing levels of disease activity tend to correlate with worsening SF36 subscale scores, especially physical functioning, however, the correlations are not tight, indicating that, in fact, the disease activity measure, the damage measure, and the SF36 or the health status measure are, in fact, measuring different domains of health-related quality of life or of health status.

I think it will be important to see what data from randomized clinical trials that incorporate the SF36 show us.

[Slide.]

Now, we have talked about the confounding issue of fatigue. We all feel that at least to some extent fatigue is part and parcel of the condition of active lupus. We

also know that there is concomitant fibromyalgia, and it has become quite controversial to some extent as to what we are measuring - are we measuring the disease itself or an associated condition?

Now, the Krupp fatigue severity scale has been utilized in lupus patients, and it has been shown that the fatigue that is present in patients with lupus is not secondarily to psychological stress alone.

Questions about fatigue are included in the SLAM and the revised SLAM. We know that fatigue can significantly impact SF36 and other patient assess measures there is no question about it. So, then we have to ask are we looking at disease-associated fatigue or are we looking at fibromyalgia, in fact, can we look at fatigue separately from manifestations of lupus.

[Slide.]

In my assessment, with my patients and a review of the literature, I think in general we can't really clarify this. There is a varying association of fibromyalgia in different cohort studies. Petri has shown that as many as 30 percent of her patients may have fibromyalgia.

Gladman and group have published that it may be about 20 percent. Drs. Gordon and Isenberg have recently looked at their patient populations in Birmingham and London, and say that it ranged somewhere between 3 and 5

percent.

Now, we have talked about the difference in fibromyalgia across the Atlantic, on this side of the Atlantic, but I think the question really may have to do with the fact that we do have very heterogeneous patient populations.

However, fatigue can be accurately measured, it can be accurately assessed over time for change, even if it does influence patient assessments, and I think it is important to note that it was just recently published in this month's Journal of Rheumatology from the group in Israel that the people with the worsening patient assessment were those with primary fibromyalgia.

The patients that had intermediate levels were those patients with lupus and associated fibromyalgia, and the ones with the least significant decrement were the patients with lupus alone. So, perhaps what we need to do at least is to look at the question of fatigue and fibromyalgia and stratify for its presence or absence at the beginning of the trial.

I don't think that we should try to force ourselves to exclude patients with associated fibromyalgia from our randomized controlled trials because I think it would be very difficult, and we would then define a patient population that will not be referable to say labeling of a

product in the real world.

[Slide.]

Disability really in lupus encompasses all domains of health status. We know that fatigue is important, we know depression is important, but we also know they are quantifiable, and what we have really learned is that disease activity, damage, and health status, as I said before, are different domains and they do not closely correlate, although they will vary together to some extent.

[Slide.]

Transition questions are obviously very important. Patient and physician assessments of disease activity don't necessarily agree, nor necessarily does their assessment of their global health agree. In fact, as much as 51 percent of patients will disagree with their doctors about how their lupus is doing, as published by Arenow and group.

I think what we realize is that psychosocial stresses will parallel patients' perceptions of their illness severity, and the physicians' evaluations, when we are looking at our patients, we are thinking of in the context of disease activity, thinking in the context of damage. We are often thinking in the context of those things that are occurring to the patient which may not, at the present time, be symptomatic, the prime example of that being renal disease. The patient really doesn't experience

renal disease until they develop end-stage renal disease.

So, we have to understand how to ask the question and what are we asking - are we asking about global health, are we asking about disease activity, are we asking about disease severity, and understanding that disease activity, disease severity do not necessarily correlate.

[Slide.]

Now, OMERACT 4 occurred in Cancun last April, and the vast majority of the people in this room, I am happy to see were at OMERACT 4, and we tried to discuss these issues in the sense of what would be the important domains to be assessed in clinical trials in lupus, be they randomized clinical trials or be they longitudinal observational studies, and we agreed that we needed to disagree, that we couldn't necessarily assign a specific disease instrument or a specific outcome measure to the domains that we wanted, instead, to reach consensus on the domains.

[Slide.]

The discussion document that I just referred to is now in the Journal of Rheumatology prepared by this group of authors, and there was a great deal of feedback to this discussion document by the participants at OMERACT.

[Slide.]

The presentations were not dissimilar to the ones that you have heard here today, and after that, the module

committee made the following recommendations to the six groups that were to sit down and use group consensus process to develop recommendations.

[Slide.]

So, we proposed to the groups that disease activity, damage, health status, toxicity, and economic costs, health utilities, be included as domains in every randomized controlled trial or longitudinal observational study.

[Slide.]

However, we also asked that the groups consider all of these other domains as possible important ones to include - death, disability, disease severity, fatigue, fibromyalgia, hypertension, psychosocial measures, serologies, working status, and global assessments by both patient and physician.

[Slide.]

Interestingly, the nominal group process result was very similar in all six groups. Now, what occurred in the nominal group process was that the groups got together and discussed in great detail publicly, among themselves, what of these particular domains they felt were important.

Then, the voting was done on a confidential basis where everyone scored the most important domain to them on a scale of 100, and the voting was then taken back

confidentially, looked at, and the three groups with the randomized controlled trials were then meaned, and the three groups for the longitudinal observational studies were meaned.

This is the mean vote for the randomized, controlled trials. Disease activity was first, health-related quality of life was second, damage was third, and toxicity was fourth, and M.D. global, economics, and patient globals were all very high out there, but did not meet the cutoff of 10.

[Slide.]

When we looked at the nominal group process for the longitudinal observational studies, we now see disease activity was first, damage was second, health-related quality of life or health status was third, and toxicity was fourth. Again, we had patient globals right here, and then comorbidities and the others were all scored very low.

[Slide.]

This was a bit of a surprise to us, but, in fact, we had a very nice consensus, and we ratified this in the general assembly meeting the next day. So, that in fact, agreement about these domains, that they should be assessed in virtually any kind of clinical trial with systemic lupus was extremely high, 85 percent yes for RCTs, 83 percent yes for longitudinal observational studies, 13 to 15 percent

no's, and a couple of percentages here who did not assess it.

[Slide.]

So, I think that we have discussed what the domains are. We can argue about which instruments would be the best ones to reflect changes in those domains, but ultimately, we do need to develop a responder index, and that responder index or that responder analysis is going to have to take into account these different domains, and it therefore is probably going to have to find a way to score multiple instruments simultaneously.

I think that is probably the only way we are going to find a primary outcome measurement in a disease as heterogeneous as lupus that will, in fact, become useful to us.

Now, in order to do that, we really need to understand what the minimum clinically important differences are in the instruments that we are going to use that reflect the domains that we think are important to measure.

We are going to have to obviously look at these things on a per-patient basis. We knew that rheumatoid arthritis was heterogeneous, but I think lupus has made this clearest to any of us who have tried to look at changes in a patient population with a therapeutic intervention over even a short period of time.

Obviously, if we can develop some type of a responder analysis, we can then compare treatments and compare treatments across patient populations. These would become much easier to use, they would be much easier to report our data, and you can gain statistical power provided we are using the components of a responder analysis that do not vary so closely together as to maximize an effect that is otherwise not there.

I think that although there is some controversy, what we have all come up with is that we have agreement on what we need to look at in lupus. We are just not clear on how to put it together.

But, in fact, if we can do it correctly and find the right instruments that reflect those domains which seem to vary somewhat independently from each other, we will actually add statistical power, and we will be able to do studies with smaller cohorts of patients.

Thank you.

DR. ABRAMSON: Thank you very much, Dr. Strand.

Are there questions? Dr. White.

DR. WHITE: I have a question that has to do sort of with theoretical considerations.

How do you think those outcomes that you all agreed upon would be appropriate in the context of a clinical trial that would be based more on an organ-specific

approach to the therapy of lupus?

What you have described would seem to be very reasonable if one were going to have an agent that you thought globally was going to affect the disease, but is it not possible that clinical trials might focus on renal disease or CNS disease?

DR. STRAND: Good question. I think, first of all, we think we may have some products which globally affect the disease and therefore we think health status should improve, damage should not increase, disease activity should go down, but then what do we do about flares which obviously are some of a component over time when you are treating a patient, and what about when we look at the patient assessments, and they don't reflect improvement.

So, I think it is maybe questionable about whether we do have products that actually globally affect the disease and result in meaningful improvement to patients.

On the other side of it, looking at organ system manifestations, for instance, in terms of a very nice set of clinical trials that have taught us a lot about lupus nephritis, I think what is important there is that we would want to say that the nephritis had not worsened, but at the same time, the patient's health status had not worsened, or that their disease activity had not worsened, or that they had not, over time, now developed new damage in another

organ system.

I think that there what we are trying to do increasingly is look at the patient in the context of the whole person. Now, as to whether this is all doable is a whole other question, but I think it is a laudable goal and I think it is an important goal because when we ask our patients how they are doing, that is an extremely important question.

I think one of the nice things about using an instrument, such as the SF36, is we are asking patients how they are doing in a variety of ways as opposed to just how do you think you are today vis-a-vis what you have just had as a therapy.

DR. ABRAMSON: Thank you very much.

We will now move to the open public hearing. Our first presenter will be Dr. Marc Gurwith, Vice President of Development and Chief Medical Officer of Genelabs Technologies.

Open Public Hearing

DR. GURWITH: Good morning.

[Slide.]

As you know, Genelabs has been conducting a number of clinical trials, large randomized clinical trials in lupus, and these are probably among the largest, if not the largest, so far in lupus. So, we would like to share our

experience and some comments about efficacy endpoints based on these trials.

[Slide.]

Our first study, GL94-01 used steroid sparing, one of the endpoints for discussion, as the primary efficacy variable. The study was initiated in 1994 and completed two years later. The entry criteria required that the patients be steroid dependent, and this was defined as needing 10 to 30 mg of prednisone per day.

The endpoint was sustained steroid reduction to physiologic levels, and we defined this as being able to reduce the prednisone or prednisone equivalent to 7.5 mg/day or less for at least the last two months of the study.

This design required a forced titration or tapering based on patient's SLEDAI.

[Slide.]

The principal advantage of steroid sparing in that it is fairly easy to measure, and it is easy to quantitate, it is unequivocal, and it has a pretty well accepted clinical benefit.

We did encounter some problems with this as an endpoint. First, the goal of forced tapering of steroids is to maintain disease activity while lowering the steroids as much as possible, but this design, as you probably realize, interferes with the ability to see if there are actually any

changes in disease activity between groups.

The study design also requires including patients who are steroid dependent, who really need the steroids. We attempted to do this by requiring that the patients not have had a flare within the previous year or be on a steady dose of prednisone for a defined period of time.

Nevertheless, once patients were on the study, a large number, more than we expected, were able to taper their steroids, and this suggested that some patients were receiving more steroids than they really needed to keep their disease constant.

[Slide.]

However, if you tried to avoid the problem of entering the patients with the minimal required steroid dose, this really means some sort of baseline period or run-in period where you force titration, where you reduce steroids again by some type of algorithm.

This kind of design may be fairly difficult in lupus. It is certainly much harder than it is in a trial for hypertension where you can fairly easily reduce the medication until the blood pressure comes up. You may not want to do that in lupus and get a serious flare, and also it may take a much longer period.

It is also difficult to devise a schedule for increasing steroids once you start tapering them and the

patient flares or has an increase in disease activity. We found, first of all, that the physicians varied in how they increase steroids at the time of a flare, and secondly, that some patients had a very large increase in their steroids, and that made it very difficult for them to taper back down to the desired physiological level of steroids.

[Slide.]

Despite the problems with this endpoint, we were able to show some differences between active treatment and placebo. This slide summarizes the results of that trial. I am sure many of you have seen it before.

On the left I have the responder results for all patients. Again, a responder is a patient who had sustained reduction of steroids. The blue bar is placebo, and there the rate was 41 percent, and it increased to about 55 percent in the 200 mg/DHEA group, and the 100 mg group is in between.

On the right is the responder rates in the subgroup defined before the blind was broken of patients with more active disease, and those were patients with the SLEDAI of greater than 2 at baseline. There, you can see an even more substantial difference in steroid reduction of 29 percent for the placebo group to 51 percent for the treatment group.

Not shown on this slide is the steroid reduction

rate or responder rate for the patients with inactive disease, patients whose baseline SLEDAI was 2 or less, and there the rates were almost identical and substantially higher, over 60 percent for all three groups and again suggesting that a number of patients who had inactive disease were getting more prednisone than they needed to maintain disease activity.

[Slide.]

Our second trial, which is an even larger double-blind, placebo controlled trial, was initiated in 1996. It is still ongoing and we anticipate completing it late this spring. It has enrolled 382 patients. Since it is double-blind and still ongoing, our observations are based on just looking at individual patient data and blinded data, so the blind has not been broken.

The goal in this study, in contrast to the previous one, it is more straightforward, and we are just trying to look at changes in disease activity and comparing placebo with active treatment.

So, we do attempt to keep the patients' lupus medications, particularly the prednisone, constant over the one year. The endpoint in this study is a by-patient responder definition.

This definition was developed after a lot of discussion in collaboration with many of our investigators

and with the FDA, but it still should be looked upon as a first-generation product.

I am sure there will be better responder definitions following this study and as other definitions are developed, but in our definition, the patient had to have an improvement or stabilization, and that was defined as the mean of all on-study visits had to be improved that is less than the mean of the baseline visits for each of these four instruments - the SLAM, SLEDAI, patient VAS, visual analog scale, and the KFSS.

So, to be a responder, the patient had to be stable or improved in all four of those instruments. In addition, we required that there not be significant clinical deterioration, but that was a fairly stringent definition.

[Slide.]

So, the advantage of this kind of endpoint is fairly obvious. It allows for integration of many different systems, and it is a by-patient analysis, which may translate to clinical benefit.

There are some disadvantages, too, of course. Not surprisingly, we found that it was very difficult to maintain a constant steroid dose over the one-year period. Frequently, the referring physician or the patient themselves would prompt an increase in steroids, and this would obviously impact or confound the responder analysis

because the increased dose of steroids would alter the SLEDAI score or other scores.

[Slide.]

Similarly, there is really no well accepted definition of what is a clinical benefit in lupus. So, a patient who met all the criteria of our responder analysis might not recognize that she had responded, that she had received some clinical benefit.

The scoring instruments aren't routinely used in the clinic, so they are not readily translated to clinical benefit, and again, in looking at monitoring the study, looking at individual patients, we see patients who, taken as a whole, clearly have had a clinical benefit and yet they fail to be responders by the analysis, and vice versa, patients being responders, but they don't really look like they improved.

Then, the scoring instruments, although validated in some ways, have not really been used or at least not used widely in these kind of clinical trials to assess change in differences between treatment groups.

[Slide.]

So, because of these problems, we also have an additional endpoint of our study, which is time to first flare. Our flare definition is very similar to the SELENA definition, a significant increase in lupus medications

especially steroids and/or a major new manifestation of lupus.

This kind of analysis also integrates many different types of disease activity, so that is an advantage, and the other advantage is that increases in steroid because of worsening are incorporated into the definition.

[Slide.]

The disadvantages are, one, although most physicians feel they can recognize a flare in a lupus patient, there is really no consensus as to what a flare is, there is no widely accept definition of flare, and then an analysis based on time to flare is really an all or none analysis.

The duration and severity of different flares are difficult to quantify, and then the flare analysis may be inordinately impacted by a single point in time or otherwise two patients, neither of whom have a flare, might differ substantially in how they did over the trial. One might improve and one might not, and yet that wouldn't be recognized in the flare analysis.

[Slide.]

So, based on these considerations, these are our thoughts. We think you should look to the rheumatoid arthritis model where the responder index, which now seems

accepted and used for evaluating drugs, was really developed retrospectively and after a large number of clinical trials, and prior to that, there wasn't any accepted primary endpoint or at least widely accepted one, and drugs were evaluated and approved based on looking at the entire pattern results in a number of efficacy variables, not just a single primary index in each trial.

Much more than rheumatoid arthritis, lupus is a waxing and waning disease, as has been mentioned several times. There is little experience with defining what clinical benefit is. So, it is going to be much more difficult to develop a meaningful single clinical benefit responder analysis. It is going to take several iterations, and that is being done right now, of course.

[Slide.]

Because of these factors, we believe that the RA model should be followed because it is premature to demand a single primary by-patient endpoint for any one single trial.

Ongoing trials, including our own, are going to help improve the definition of responder, and until then, it is going to be more important or better to again look at the pattern of results and make the assessments as was done in arthritis.

Thank you.

DR. ABRAMSON: Thank you very much.

The next presenter will be Dr. Jill Buyon, Professor of Medicine at NYU and the Hospital for Joint Diseases.

DR. BUYON: I thought I would really like to present four slides that represent true user issues.

[Slide.]

These reflect the input of several investigators - Joan Merrill, Ellen Ginzler, Michelle Petri, Ken Kalunian, Mike Belmont, and they really reflect our usage of these instruments as participants and also as consultants to some of the pharmaceutical trials that have been launched.

[Slide.]

How do you meet the needs of a multicenter trial? I think one of the most important issues that has come up from a practical point of view is educating investigators to ensure uniformity of data collection.

I really thank LaJolla Pharmaceuticals and Abbott who recognized this before their launching their trial, and so in most trials, we would suggest the provision of specific guidelines for each instrument to be used in the trial, for BILAG, for example, a glossary of terms or a user's manual, agreement on how to score the descriptors, what exactly is being measured, emphasis on attribution of signs and symptoms to lupus, use of paper patients to assess investigators' use of the instrument, and lastly, education

of the study monitors who actually are coming to look at your books.

[Slide.]

This all really came about with the first multicenter trial that we have just heard about from Dr. Gurwith, and I want to give you three examples. You take a simple instrument like SLEDAI. It seems very straightforward, and these were questions that came out of scoring the descriptors, and these were the answers.

First, people asked, "What is the definition of new or recurrent for purposes of scoring the ulcers, alopecia, and rash on the SLEDAI?"

And these were the answers given actually by Dr. Kenneth Schwartz, and you will see how they bring up several problems. He answered this question as, "New or recurrent means either the finding has never been noted before or has come and gone previously, but has recurred within the last 10 days. Lesions which have recurred outside the 10-day window should not be scored."

How do you score a rash which worsens on the SLEDAI? If it is the same rash which either extends or worsens within the same localization, it does not get scored on the SLEDAI. If it was a new rash, it does get scored.

Should rashes which are improving be scored on the SLEDAI? No, unless they have not been previously been

scored on the SLEDAI.

[Slide.]

Now, how does this pose major problems for a trial? I illustrate this in a sort of mock chi square, if you will, the advantage of SELENA SLEDAI over SLEDAI, and one of the changes we made was from descriptor adjectives new or recurrent to ongoing, and I am going to give you two examples.

Look at the SLEDAI at the qualifying visit. The patient on placebo has a new rash, 2 points, the patient with active drug has a new rash, 2 points. At the one-month visit, the patient on placebo has a worsening rash, but we already heard the definition, and now the score has to be zero, and the person on the active drug has no rash, again scores zero. Your interpretation would be that drug is equivalent to placebo.

Let's look at the SELENA SLEDAI. Qualifying visit, the patient on placebo has a new rash, 2 points, the patient on active drug has a new rash, 2 points. At the one-month visit, patient on placebo has a worsening rash, and this is again scored as 2 points. The patient on active drug has no rash, receiving a zero. Interpretation, drug superior to placebo.

[Slide.]

What about evaluation of drug efficacy? That is

important to acknowledge the heterogeneity of SLE. We have discussed that by organ system. Are we treating severe lupus, mild-moderate lupus? DHEA and cytoxan would be for very different populations of patients. So, we need two kinds of indices.

An activity index defines how ill is this patient. A responder index defines whether the patient got better and by how much. So, an objective in many trials would be to compare across the spectrum of patients, not to ask how sick is this patient, but how well did the drug work for that initial objective.

So, I will give you some examples. At qualifying visit, arthritis in six joints, SELENA, they would receive a 4. In the RIFLE, which is done only at two time points, beginning and end, it would be present.

What about thrombocytopenia, 500 per millimeter cubed? One point on SELENA SLEDAI, which does illustrate some of the inherent problems in the SLEDAI because of some weighting issues. RIFLE, present. What about catatonia, 8 points on SELENA SLEDAI; RIFLE, present.

Now, we will look at the termination visit. Arthritis in three joints, SELENA SLEDAI, 4; but in the RIFLE, this would be recognized as a partial response. Thrombocytopenia now 99,000, one point on SELENA SLEDAI; partial response.

The patient who was catatonic is now awake, but still has impoverished thought, 8 points, but a partial response.

Our conclusion would be from a pharmaceutical point of view, we have just gone from 13 to 13. In every category, there was no change, but in the RIFLE, this patient would be considered a winner.

Thank you.

DR. ABRAMSON: Thank you, Dr. Buyon.

We will now take a break until 10:40 and try and reconvene promptly at that time. Thank you.

[Recess.]

Panel Discussion

DR. ABRAMSON: We have been asked to address five questions this morning pertinent to the presentations that we heard. We have invited panelists to join us as well this morning - Dr. Paul Fortin, Ellen Ginzler, Ken Kalunian, and Jack Klippel.

What I would like to do to start off addressing these questions, which we would like to finish each of them by 12:15, is begin with our expert panel members and ask them, looking at Question 1 particularly if they would like to open up the discussion with a comment or two about Question 1: What claims would represent a clinically important benefit in SLE looking at some of the potential

claims that are listed below.

DR. FORTIN: Reading this, I first thought, and being a participant in both SLICC and OMERACT, I obviously have given some thought about that, and I think that disease activity is certainly an obvious important outcome.

I was just thinking, though, that as was discussed a little bit earlier, the problem is how to report on disease activity in a randomized, controlled trial, and this is obviously a little different than in a longitudinal observational study, and that methodology, there are some challenges. The challenges are if you use a mean change for a group, you may lose a lot of the actual impact of your intervention in a population of lupus patients where the disease is waxing and waning over time.

So, there is the issue of the duration that will be linked to how you decide to define your disease activity and the issue of the methodology. If you use a group mean, you would probably need to have a very large, consistent response in order to detect a clinically relevant outcome.

If you look at disease activity within patients, then, that is a different issue, and you may be able to identify subgroups of patients that would respond. So, that is with the disease activity.

The disease damage is much more difficult. I think it is more of a characteristic of the population you

are going to be including in your clinical trials. You would have to have either a very, very large sample size to detect no progression in disease damage, since it cannot change over time towards decreased damage, it can only go up. So, you do a very large sample size or a very long duration in follow-up since damage won't be changing for at least six months by definition if we use the SLICC as your damage index.

For health status, in my mind, it is also an important outcome since we are obviously interested in determining what will be the impact on patients. It will depend a little bit of what the focus of interest will be. I think activity and damage have been defined by physician and most likely to reflect the physician's view. The health status is really reflecting more of the patient's view.

The other questions were for the organ-specific disease, and there were some comments a little bit earlier, I think by Dr. White, about how can one instrument, if a drug intervention will be only aimed at, let's say, renal damage, how will it involve the general activity in lupus.

I would think that even though an intervention may be with primary outcome an organ-specific instrument, organ-specific outcome like renal disease, it may be very interesting to know if there is an impact on global disease activity and, generally speaking, the health-related quality

of life.

The issue of fibromyalgia and fatigue, again, there was some discussion by Dr. Strand about this issue. Specifically, clinicians, we all know that fatigue is always a very important complaint of our patients and that it may or may not be correlated with fibromyalgia, however, if we are to single that specific characteristic of lupus out, it is methodologically challenging, but it may not be alone, a stand-alone criteria unless the aim of the trial is only to improve that specific outcome.

I just wanted to comment in flares, I guess the correlate of flares, we haven't talked about remission. I think people don't like too much the word remission, but maybe response, and yet it is interesting to document per patient the number of exacerbations of flare and whether these are complete, you know, severe, mild, or moderate, but the response also needs to be defined further and also the study design of your randomized, controlled trial will have to integrate whether your primary objective is to define response or whether it is to define the number of flares or reduction in number of flares, and they are very different questions and will imply very different study designs.

DR. ABRAMSON: May I ask, as we go down the panel, just to focus things a bit among these questions, is the relative validity of disease activity measurements in a

global way in a heterogeneous disease like lupus versus the other question effective for organ-specific diseases, and how you weigh those two elements in thinking about this process.

Dr. Ginzler.

DR. GINZLER: I would like to focus my comments I think on some conceptual issues. Clearly, the drugs or the modalities that are being developed and tested now appear to be aimed at our understanding, which is ever-increasing, of the pathogenetic mechanisms in systemic lupus.

So, whether a new treatment is aimed at specific disease activity or global disease activity, and therefore specific end organ function and/or damage versus overall patient outcomes may depend to some extent upon whether that modality is aimed at changing some specific immunologic dysfunction or some specific pathogenetic mechanism, and clearly we see differences in that in terms of the agents that have been brought before us now.

I mean certainly the rationale for testing a drug like DHEA is related to the known hormonal effects of the disease. So, one might expect that there would be differences in an agent like that compared, say, to a B-cell tolerogen in terms of what we can expect, and I think those things have to be kept in mind.

The other issues which I think again are of a

conceptual nature relate to the assessment of damage and how that relates in a short term versus a long term trial.

I think that there are both short and long term damages. I mean we have already said that it has to be irreversible, and a proxy for irreversible is that it is present for at least six months, but there are elements of damage which do occur by six months or a year which might be within the realm of the time that a specific trial is ongoing as compared to damage which may accrue over five or 10 years, and Dafna did show that that accrual of damage continues over that long period of time.

Damage is very much related to quality of life, and there are again short term quality of life issues versus long term quality of life issues, and for some of our patients, those may be quite individualized.

I mean one patient may find that spending the rest of their life on hemodialysis is really not a bad thing as long as they maintain their attractive appearance, and, in fact, I have had patients who have refused treatment with steroids and other agents.

One in particular, a beautiful, young teenager whose sister was on dialysis, and she said, well, I know what dialysis is, I can live with this, I refuse treatment, and, in fact, her kidneys failed and she went on dialysis and is doing extremely well with it. For someone else, that

quality of life might be intolerable.

So, some of these issues have to be considered, but they don't alter the actual definitions of whether a particular agent satisfies the criteria for response. So, those things have to be thought about as we do this.

Finally, since we are learning more about pathogenetic mechanisms, some of the things which we might attribute as damage -- and remember in the SLICC damage instrument we don't count attribution -- it is merely damage that occurs after the diagnosis of lupus, but not separated out whether it is due to lupus per se, toxicities of treatment, or to intercurrent illnesses.

Remember that there are, as I said, short term, as well as long term, effects, and a perfect example is the issue of atherosclerosis. For many years it seemed just unquestioned that corticosteroids caused accelerated atherosclerosis.

I am happy to say that I was one of the early people to say that I didn't think that that correlation existed, and, in fact, I think it is probably a proxy for severe disease, and now that we are learning more and more about the interrelationships of the immunopathogenesis of atherosclerosis, and the abnormalities that occur in the immune system in terms of endothelial cell function in lupus, I think that is becoming much more clear.

We are not going to have the luxury in almost any trial of looking at damage, say, in development of atherosclerosis over the life of most trials.

I mean if a trial takes six months or a year or even two years, the likelihood of a statistically significant proportion of patients developing clinical manifestations of atherosclerosis during that period of time is very small, but over a long period of time, we will have to look at those issues, just like we have looked at how often and in what population infertility develops after cyclophosphamide treatment.

So, I think those are all issues that we need to keep in mind when we decide how we are going to use these measures, as well as specifically what the measures are.

DR. ABRAMSON: Dr. Kalunian, which of these things are potential claims that drug companies should go after?

DR. KALUNIAN: I think all of these things are potential claims really, and I think that they need to be measured concurrently.

For example, if you are trying to understand whether a novel agent affects disease activity, I think you have to understand what impact any incurring damage has had on the ability for activity to change.

For example, if you have damage in a target organ, that organ may not be able to change, and you need to know

that in terms of assessing a drug's effect on the individual person.

I think that it is important to sort of ferret out the effect of a novel therapeutic agent on fibromyalgia or fatigue to understand what effect you are having on other aspects of that disease.

So, it is important to understand where you are targeting, what the drug is actually targeting, whether it is drug targeting global effectors of health quality under which fibromyalgia and fatigue may fall, or other aspects.

I think that when you get into the issue of organ-specific diseases, you really need to assess your measurements of disease activity, damage, quality of life, et cetera, on instruments that actually can focus on those target areas.

I think Dr. White brought this point up earlier about which of the different disease activity instruments are good for different aspects of disease, and as somebody who has actually used these instruments extensively in clinical trials, as well as in following my patients, I think, for example, the SLAM instrument is very good at looking at patients who have relatively lesser degrees of disease activity, whereas, the BILAG and the SLEDAI tend to address patients who have a more severe, if you will, spectrum of disease activity.

So, I think all of these potential claims are necessary, but you need to also look at the ability to measure them and you need to be practical, and they also have to address the issue of response of the target organ depending on whether you are looking at a target organ or more global disease activity or spectrum of disease.

DR. ABRAMSON: Thank you.

Dr. Klippel, are each of these stand-alone potential claims?

DR. KLIPPEL: No, I think this list really only has two items that are important for claims. One is the effect of a drug on organ-specific manifestations, and the other is on disease activity, and on those two, I believe that it is the organ-specific one that is by far the most important.

I believe that drug development is likely to be biologically based, so that one will go into the trial not as an empiric trial, but with some understand of what the drug or biologic is likely to do, and that will help with the targeting of how the drug should actually be tested, so that the claim likely will come forward based on the effect of the drug's known biologic effect on an organ-specific manifestation.

So, I think that far and away, in my opinion, is the most important on the list. I do believe that it is

important to measure disease activity with one caveat. It seems to me that that almost implies that one is using an empiric-based notion for drug testing.

You are testing the effect on lupus activity, and I don't know how then the product would be marketed, because a practitioner then wouldn't know exactly how to use the drug when he or she has a patient in front of them with the exception, if the disease activity were to go to zero, that is, I think we are hunting at this stage.

One of the things that distinguishes lupus from other rheumatic diseases is that we do have therapies which affect this disease, and that the next step forward needs to be to identify therapies which actually induce a state of quiescence, if you use Michelle's term, or state of remission, and for that purpose, I think you need to measure disease activity and ask the question does it go to zero.

DR. ABRAMSON: Thank you.

I would like to open it up now to members of the panel and this morning's presenters to look at these potential claims. Obviously, in Question 2, we will talk about what is the best way to measure these outcomes, but what do people think about each of these as claims?

Dr. Lovell.

DR. LOVELL: I have a question for Jack Klippel. Why don't you think damage is an appropriate claim for

lupus?

DR. KLIPPEL: I think the ability to measure damage is terribly important for people who do longitudinal studies, who are trying to understand the clinical course of the disease. Unless there is a major change in drug testing, it doesn't strike me that clinical trials of drugs are going to go on long enough to actually see an impact on damage.

DR. LOVELL: How long do you think that has to be, and isn't that somewhat equivalent to what we are asking people to do in RA trials in terms of prevention of radiographic damage, and that sort of thing?

DR. KLIPPEL: It is much simpler in RA. You are talking about one organ system in which you are measuring damage, and, of course, the current ability to assess damage is very complicated.

There is multiple systems involved, and Dafna is the better one to answer this, but I think you are talking about a year or more, and that may be an add-on to all clinical trials, that you may want to stipulate to ask the question about damage, but I think it is quite unlikely that anyone is going to be able to show with a drug that you are truly minimizing damage or altering the course of it.

DR. ABRAMSON: Dr. Silverman.

DR. SILVERMAN: When I look at these, it strikes

me that they are a dichotomy, and I think Dr. Ginzler hit on it, the difference between quality of life and fibromyalgia versus the actual disease activity, and she mentioned the patient who went on to dialysis, I don't think anybody at this table would think that is a good outcome.

We all know that steroids certainly alter the quality of life for the worst, but we all use them, so I think that when we are measuring quality of life and effect on fibromyalgia, they are almost different endpoints, and one has to be careful if you are going to include them in an endpoint, especially off a short term trial, under a year, that if we put too much emphasis on things like quality of life, I hate to say it, and effect on fibromyalgia, i.e., steroids, can certainly make it worse, that we may in fact weight it to an outcome that we actually don't want, end organ damage, so I think we have to really be careful and maybe go for two separate claims.

DR. ABRAMSON: Dr. White.

DR. WHITE: I would like to echo Jack's comments. From my standpoint, it would seem to me to be particularly important to have the flexibility to have some organ-specific claim. I think there are probably more objective measures of an individual function of an organ in some of the indices that we have developed.

I think that it would be easier to define a

homogenous group among a heterogeneous population, and I think that it would be critically important to allow that claim to be established without necessarily requiring that you show global disease modification as assessed by damage or activity.

I think it would be very important to ask such sponsors to collect those data. It may give them a better claim in the long run, but I don't think that that should be required.

I also have some concerns about the damage claim. Given what we were told about the damage index, it seemed to me that a lot of the scoring depended upon drug toxicity or at least some of the scoring, as was clearly acknowledged.

Therefore, that seems to me to be a very difficult experimental design, and I just have some concern that that might be a difficult claim to actually meet given we have an instrument, but we also know there are some limitations.

DR. ABRAMSON: Dr. Gladman.

DR. GLADMAN: I think that there are actually three separate claims, and it was pointed out in the fact that three different people got asked to discuss things, and Vibeke actually put it all together.

The first claim that I think is important in a drug trial is it needs to reduce disease activity, and whether it reduces disease activity globally or whether it

reduces disease activity in an individual organ will depend, as Jack Klippel said, on the nature of the drug, the biology of the drug, and the expectation from the drug.

But regardless of which organ we are talking about, we cannot use the organ outside of the patient. The patient has disease activity usually in lupus in more than one system, and therefore, a global measure of disease activity, and if we are talking about claims, an improvement in the global measure of disease activity is just as relevant as the improvement in the particular organ, because to treat the one organ and wreck several others at the same time, is not a winning situation either.

Now, Jack was talking about damage as if it is a totally different approach, however, in all the drug trials of lupus nephritis, one of the outcomes has always been going on to renal failure. Well, that is damage. That doesn't say anything about the activity of the kidney, it is a failure of that kidney.

So, inherently, people have already incorporated an organ-specific damage into a number of observations and outcomes that have been incorporated into drug trials.

What the damage index provides is the ability to look at individual, as well as global, damage in a patient, because within the damage index, there are 12 different organ systems, and one can look at them individually related

to the organ that we are interested in, in terms of activity and in terms of damage.

But I think it is important, though, that the activity is separate from the damage, is separate from the function of the patients, because we have shown, and other people have shown, that the patient's perception of their quality of life and health-related function is a distinct domain.

There may be some relationship to disease activity, but it is weak, and it is a distinct domain. People feel how they feel for their own reasons, not necessarily related to the disease process itself, and this has been shown for lupus, it has been shown for rheumatoid arthritis, it has been shown for diabetes, AIDS, whatever disease you want to look at.

It is an important outcome, though, because at the same time that we are trying to improve the disease process for the patient, we are not trying to make them more uncomfortable or less functional, so that one does not necessarily expect that the patient function would get better, because there are so many other contributors, but we certainly do not want to make it worse.

I think that claims can be made about decreasing disease activity, preventing worsening of disease activity. For example, we recently presented at the ACR meeting an

approach to the SLEDAI which is based on data that we collected that suggested a remission is a SLEDAI of zero. An improvement is a decrease of SLEDAI by more than three points. A worsening or flare is an increase of the SLEDAI by four or more.

We have the concept that if the SLEDAI had increased or decreased by less than three, the patient is either persistently active or unchanged, and therefore one would not claim that a drug has actually improved the patient if the change in the SLEDAI was so minimal.

So, there are ways to approach the assessment of disease activity. Now, if you take the BILAG score, you actually can address the individual organ. If you want to choose kidney, you choose the renal system. If you want to choose hematologic, you use the hematological system, respiratory, whatever.

Then, you can actually measure an improvement or worsening by going up and down the scale of A to B, so those are available.

DR. ABRAMSON: Thank you.

Dr. Strand and then Dr. Isenberg.

DR. STRAND: Well, I think the point here that I feel is confused, if you would consider each of these as separate claims, is that you could treat the disease and throw the baby out with the bath water, and the patient

could be worse.

I think we don't want do that however we want to look at it. Even when we look at renal disease, 81 percent, 70 percent of patients have other manifestations, other organ system manifestations, so it doesn't seem to me that we are actually looking at a very accurate reflection of the disease process if we limit ourselves to an organ-specific manifestation.

We could still make one of those particular outcomes primary until such time as we can put everything together into a responder index, but I just cannot see us looking at disease activity without noticing whether damage has increased or not, and without noticing whether patients at least have the same health-related quality of life and health status or at least some improvement of it.

From that point of view, it seems to me that that is why we are talking about at least those three domains in each clinical trial. From that point of view, we ought to consider all three domains within a claim as well.

DR. ABRAMSON: We would agree that a drug does not have to achieve all A through E, that a drug can achieve a specific organ effect obviously and still be valuable and approved so long as it wasn't doing negative things to other organs.

Dr. Isenberg.

DR. ISENBERG: I think Jack Klippel is right in that one might want to consider whether there were one or two prior claims or particular claims with respect to activity in particular and to individual organs, and I think Dafna has argued the case for why the BILAG system provides us with the ability to do that.

But I would certainly want to push the issue of damage. After all, it is exactly 50 years since the first human being was given a form of ACTH for their rheumatoid arthritis. I think it is very unlikely that those who gave that patient the ACTH expected that that patient many years later might develop cataracts, might develop proximal myopathy.

The things about lupus is that you have to expect the unexpected, and I think the damage index allows us to do that.

DR. ABRAMSON: Any other questions? Dr. Lovell.

DR. LOVELL: I have a comment. I think, Vibeke, that if you take the NIH cytoxan trials for kidney disease as a prototype, then, I think there can be a very strong case made for looking at organ-specific indications. I mean I think that was a major breakthrough in our understanding of treatment of lupus, and that was a very organ-specific sort of approach.

So, I would think that historical precedent sets

the importance of looking at organ-specific treatment for lupus disease.

DR. STRAND: I wasn't disagreeing. I was simply saying that if you are going to look at the organ-specific manifestation and make that the primary outcome, for which I have no problem, it is important to include as secondaries that these other manifestations do not worsen.

DR. ABRAMSON: Dr. Fernandez-Madrid, do you have a comment or a question, and then we will move on to the next question.

DR. FERNANDEZ-MADRID: I think there is a definite relationship between disease activity, persistent disease activity will lead to organ damage and organ-specific disease, and I think there is a relationship among them.

I think that what we want to prevent is organ damage and organ-specific disease, but I would like to make a comment on something that was said this morning in reference to that the mortality is not an outcome measure in lupus anymore.

I think we know that survival has improved. After 10 years, maybe 80 percent of these patients survive and have a normal life, but 20 percent may die. Particularly in vulnerable populations like young African-Americans, this is maybe a more important problem.

So, I would not eliminate this issue from the

possible claims.

DR. ABRAMSON: Thank you.

We are going to move on to the second question. I guess the first question is agreeing upon what outcomes we are looking for, and the second question is how best to measure those outcomes. We heard a variety of indices that are available this morning.

We can open the discussion. Which of the following represent potential endpoints? We will begin I guess with the responder indices and talk about the pros and cons of perhaps what we heard this morning.

Someone raised the question of can we have an ACR 20 equivalent for lupus like we have in RA. Do any of the indices lend themselves to that kind of assessment, comparative assessment of drugs?

DR. FORTIN: I would like to even broaden the issue if I can. Could an indices include a disease activity or disease damage and have related quality of life? Obviously, we are not only measuring joint count, which has a very finite variability, but we have notable composite indices that we would then pool, so that we increase the challenge, but we can think of a responder index like the RIFLE, which is one single instrument, but couldn't we also think of a composite of these three other domains?

DR. ABRAMSON: I would like to turn to the panel.

Dr. Liang has been very quiet. I am curious if Matt wants to begin this discussion on this side of the panel.

DR. LIANG: Which one do you want me to be allowed on? I have really enjoyed the discussion. I think it is a little too soon to close on this issue. I have heard response index described. That is combining things.

I heard from Vibeke that the implication was that we might try to mesh quality of life with fatigue, with one of these disease activity indices, and I think that perhaps it would be more strategic and more useful, and it is probably not so important that we be right, but be consistent in our terminology, that we define on an organ-by-organ basis what is a clinically meaningful response, and the most physiologic and convenient way that we can do that, and then later on perhaps to see if we can improve statistical power by combining things as diverse as fatigue, quality of life, and disease activity.

So, I would be inclined to sort of crawl before we walk, before we run.

DR. ABRAMSON: David.

DR. FELSON: I have a lot of related sort of concerns and questions. I am struggling a lot with how incredibly messy this is, messy from almost every measurement perspective I can think of, and I wanted to offer a number of other than just use the deconstructing

term messy, offer a number of thoughts and suggestions that might help with the messiness.

One is the notion that one might increase power to detect change by doing what Dafna started to suggest, that the Canadians had begun to do, which is characterizing an individual patient in a trial by several levels of transition rather than just one.

What the ACR 20 is an example of is improvement or no improvement, and that happens to work pretty well in rheumatoid arthritis because it so happens there is a bimodal distribution of change over time. That is what we sort of found.

But it probably is not likely to be the case in SLE, and I think for an individual patient, one might characterize their changeover time as worsening, as what Michelle characterized as a mild flare or a severe flare or improvement, so there might be four states over time that an individual traverses - no change, improvement, mild flare, severe flare.

That gives you a state of four, and at multiple time points. So, it is not time to flare. What one could do is then use time information to get more information. I mean the more information one can get on both an individual in terms of their severity or improvement -- and that is another issue, by the way, transition versus state, and I

think it is really incredibly messy -- and using multiple time points to get that measure starts to give you efficiencies in doing trials that really we all need here.

Getting away from the measurement, the incredible measurement complexity to just say, look, from a clinical trials design perspective, what would help. What would help would be probably to define each person's outcome nondichotomously and each person's outcome at each time point, so that one could then use multiple time points for each person to get information. That would make it more efficient.

Then, there are two measurement issues which relate to disease activity measure, which I am mind-boggled at their complexity. Matt just introduced another issue that is also bothersome, which is whether you need to score each organ, get an organ-specific severity, and then sort of summarize them.

But all of these instruments seem to have been transition instruments, so at every point in time, patients are assessed as to whether they have developed a new manifestation, a changed manifestation or an improved manifestation.

So, essentially, what you are doing is measuring the change of change in these trials, which we debated this in the rheumatoid arthritis efforts we went through. There

is a considerable amount of data from pain studies as to whether transition versus state instruments are more or less sensitive, and the data is somewhat conflicting, but the two correlate extraordinarily highly, by the way, transition instruments are very messy to use in trials.

For one thing, people don't remember the baseline, so the further you go from the baseline, the more inaccurate and noisy the transition measure becomes.

The instruments that I saw don't have the base state defined well, and that introduces variability, so what happens is you say yes, in the last month, has the patient developed pleuritic pain, compared to what baseline time, and does the patient remember the baseline time, and was the patient examined at the baseline time.

I don't know how all these instruments are constructed to address that concern, but one would have to think about that concern.

The last extraordinarily messy issue that I wouldn't know how to suggest or even think about dealing with is whether to incorporate other treatments as measures of response.

Generally speaking, in RA and OA, that is something that we have strongly discouraged. So, you don't usually use as a primary measure of efficacy in an OA trial the use of acetaminophen.

The sponsor who presented data on steroid use response using that primary efficacy measure very nicely pointed out all the pros and cons of that, and I realize in lupus, it is not so straightforward because you have got prednisone widely used, and it really could determine disease activity, et cetera, but I am not sure exactly what to suggest here, but I would strongly suggest that you try to develop, or one, try to develop or adopt a disease activity measure that doesn't weight heavily the dose of steroid use because that in itself introduces physician treatment variability, and if you move from SLEDAI definitions to steroid response, it introduces a whole other set of issues that he commented on, ceiling and floor effects of steroid doses and whether steroid doses are needed at certain levels. It is just very messy.

DR. ABRAMSON: Just to pick up on that to try and get some clarity as we go forward, I just want to pose a couple of questions, and, Ellen, if you could hold your comment maybe in the context of this.

Because so much work has gone into the various indices that we have, and they clearly are indices in evolution and trying to adapt to some of the changing drugs available, et cetera, if we looked at these top four items that are presented to us as endpoints - responder indices, flares, disease activity, and damage, I want to see where we

have consensus and where we have disagreement among those four things among the experts who have spent years looking at these issues carefully.

So, damage, we have a SLICC instrument that I guess is the only one on the block, but those of you who have spent time doing this, could you look at these and talk about damage first, where do we have consensus among these four items and where do we need to develop a finer instrument that addresses some of the things that David Felson was raising? Ellen.

DR. GINZLER: Well, I think obviously, you have to move drug toxicity up into the issue of damage. I mean it is the first kind of damage.

DR. ABRAMSON: I want to focus on where we have instruments, just damage and above, where do we have an instrument?

DR. GINZLER: In fact, we have a damage instrument which fortunately or unfortunately has been agreed upon, so it is the only one where there aren't competing instruments. I mean there is an accepted instrument.

As I mentioned before, I think that the biggest problem with it is short term versus long term damage in terms of which manifestations we might expect to see occur during the lifetime of a clinical trial as opposed to the much broader experience of a longitudinal observational

study. But the instrument itself has been validated and it has been well accepted, so I think that that is not a problem.

We have a number of disease activity scales. A number of them have been not only validated, but compared, and I echo the comments of people who have said that you may select a particular disease activity index based upon what kind of agent you are testing and what outcome you are looking for.

So, again, I don't think there is a need to develop a new disease activity index, and I think that the ones we have are available and useful.

The definition of flare has not been codified. I mean there have been a number of studies that the Toronto group, the Hopkins group have looked at that. There have been definitions of flare. They haven't been tested in other cohorts. They are based primarily on numbers, not on conceptual changes in disease.

I think those of us certainly in SLICC would agree that flare is not well defined, and we do not have a good instrument that is acceptable for all studies. Likewise, the issue of responder indices, I mean RIFLE is a work in progress.

It has an enormous number of problems and issues. It is extremely complex at this point, and I think there is

no doubt that from the point of view of statistical power, if we are going to proceed with that as a potential instrument, a lot of work needs yet to be done.

Some of the items may need to be condensed. We need and we have been looking at this, a head to head comparison with BILAG, which was not developed as a responder index, but certainly has many of the features which could be used as one, and certainly many of the same concepts that are in RIFLE.

So, I think it is a work in progress, but we don't have an instrument at this point for either flare and concomitantly remission, nor do we have one for responder index.

DR. ABRAMSON: Thank you. Dr. Isenberg.

DR. ISENBERG: I would like to try to help tidy up the mess. I think, briefly, there is one agreed damage index. Most people accept SF36 as the health perception index. As far as activity indices are concerned, although I think David is right to be concerned about, shall we say, some of the more precise formulations of how these things came to be developed, and so forth.

Practically everybody agrees that we look at activity over the previous month. As I have shown you, a number of different groups have compared the global score indices and found them to be remarkably comparable, and

these are studies done in the States, done in the UK, and done in Europe, and furthermore, we have from Stefano Bombardiere a computer program which, from the same database, will give you a SLAM, will give you a SLEDAI, will give you a BILAG, will give a SIS, you know, that already exists.

If you want something a little more sophisticated, then, you could either use the BILAG or, you could, if it is perhaps modified a little, use the RIFLE.

I think although there are legitimate concerns about the minutia of precisely how we have come to define things, and we could always work on improving that, I think, by and large, there is much more clarity here than perhaps one needs to be concerned about.

DR. ABRAMSON: Dr. Strand.

DR. STRAND: I would like to echo that comment from David. I think the important thing here is that we are trying to look at an intervention that is now going to result in some type of improvement in manifestations of lupus.

Now, disease activity indices, in and of themselves, were never designed to be endpoints, so I think in the context of that, they cannot be a stand-alone endpoint by any means, but we certainly can use any one of them.

Now, if we look at those scores, if we look at those in the context of what the health status or health related quality of life of the patient is, and then we ensure that damage has not increased, then, we have some kind of a feeling that at least we are not, shall we say, sliding backwards.

In terms of flares, a flare may happen in the context of a treatment because a treatment has not had such a rapid onset of effect or whatever else, or because we already know this is a disease that flares and remits, and flares and remits, or shall we say, improves and worsens.

Really, aren't we better to look at an area under the curve analysis and over time analysis, as David is suggesting, where we don't necessarily have to have an agreed definition of flare so long as we can say that over the majority of period of time that we have looked at a patient when a certain intervention has been made, they are at least the same or hopefully improved to what they were before.

I think we can move on to responder indices over time. We are not ready for them. We don't have the data from randomized, controlled trials in lupus that we had from RA, so we could define a responder analysis and then, in fact, validate it.

We are not there yet, and I think that if we try

to do something too quickly, as Matt has said, we will throw the baby out with the bath water, but the other part of it is most of our instruments are developed from observational studies, and we are going to have to use them in RCTs before we can be sure that they will do everything we want them to do.

I think that is our most recent experience with these disease activity indices.

DR. ABRAMSON: If we all look at this list provided to us, are there items that are not potential endpoints? David, were you suggesting that steroid sparing was not a potential endpoint, for example?

DR. FELSON: I would struggle with that, I think. I think it is a secondary endpoint for sure. It is obviously of great importance, but I think I would be inclined to say it isn't a primary one.

DR. ABRAMSON: Dr. Silverman.

DR. SILVERMAN: I can understand David's point from a methodological point of view, but I can't understand it from a clinical point of view. Steroid has to be, I mean it is the number one use. I can design a trial, a six-month trial that every single drug will work if I set my steroid dose, it will work in 90 percent of the patients.

So, it is so crucial, and I understand that it muddies the water, it makes it messy from a methodological

point of view, but I don't see how we can have a trial that steroids aren't controlled in or at least taken into consideration as a major endpoint.

DR. ABRAMSON: Dr. Fernandez-Madrid.

DR. FERNANDEZ-MADRID: It seems to me that what muddies the water is the excessive use of corticosteroids in clinical practice, and when you see, and everybody sees these patients, they are taking doses of steroids which are far excessive of the needs to control activity.

So, I think a steroid-sparing effect in this context will have to be preceded by a critical appraisal of the dose of steroids in these patients prior to a trial.

DR. ABRAMSON: Any other comments on this piece? We are going to come back to some of these in Question 4, I suspect.

Let's move on to Question 3. What is the appropriate duration of clinical trials in SLE?

Dr. Kalunian.

DR. KALUNIAN: I think that really depends on what drug you are studying and what organ you are targeting.

DR. ABRAMSON: Dr. Klippel.

DR. KLIPPEL: Twelve months.

DR. ABRAMSON: I thought your data took five years to see that steroids didn't work.

DR. KLIPPEL: The thing that Michelle nicely

pointed out is that this is a disease that is so variable over time that the danger here is that the trial is going to be so short, that that is going to override your ability to interpret anything, and the practicality is that you have to identify something between a matter of hours or days and the patient's life and what society can afford, and 12 months is right there.

DR. ABRAMSON: So, that is a minimum.

Other suggestions? Dr. Harris.

DR. HARRIS: Certainly, 12 months seems rational depending on what you are measuring, because certainly if you are looking at disease activity, for instance, I can imagine that there are agents wherein one would have decreased disease activity in a period less than 12 months. On the other hand, if you are looking at organ-specific damage, I would, in fact, argue 12 months or even more.

So, from my perspective, it depends on what parameters you are measuring.

DR. ABRAMSON: Dr. Moreland.

DR. MORELAND: As I see the relapsing-remitting issue here, has the group given any thought as to whether you should have a run-in period, standard run-in period of a month, or should you just randomize everyone, so that there is standardization among the clinical trials?

DR. ABRAMSON: Dr. Petri.

DR. PETRI: I suggested that the ideal patient to randomize in a lupus clinical trial is someone with chronic activity. I don't think a run-in, where we reduce prednisone is ethical in a patient who has chronic activity, but if we pick the patients with the chronic activity, we are going to find it, I think, easier to evaluate them, and it would allow us, I think, to have trials that are shorter than 12 months.

So, to give an example, I would only enroll patients who have active discoid lesions if I am doing a cutaneous lupus trial, because those lesions should hopefully respond to an effective treatment, let's say, in three months, why would I have to wait 12 months.

DR. ABRAMSON: Dr. Liang.

DR. LIANG: I was just going to add another "it depends." I think the goal may be to find a remittive agent or it may be a maintenance agent, and I would be happy to do a one-month trial in an induction study where I am looking to wake up people in coma, whereas, I might be more interested in the long term issue in terms of whether a drug can maintain a person in a certain state.

So, I think there are a lot of factors that go into the duration.

DR. ABRAMSON: Dr. Ginzler.

DR. GINZLER: I echo that completely. I think

there are two things that you can look at, and obviously, it depends upon the feature. I mean if you have something like skin disease, you can look at time to response and then you can look at area under the curve over a given period of time like a year.

I mean we have all seen rebound, so if someone responds very quickly, but then a month later, they are sick again, that might be a great drug if there is a second drug that is useful for maintenance.

I think these are things that we need to consider, and as Matt says, if someone is comatose, or if their serum creatinine is rising very quickly, I mean you don't have the luxury of observing them for a month to determine what their premorbid pattern is.

So, I think that we can measure both time to response and area under the curve, which will show us the pattern over time as they remain on that agent with the modifications based upon what the particular manifestation is.

DR. ABRAMSON: Shall we move on? Let's go on to Question 4. What parameters should be measured in all clinical trials in SLE? I imagine this brings us back in my mind to the issues under Question 2.

Would anyone like to start this discussion?

Dr. Gladman.

DR. GLADMAN: I think it is clear from everything that has been said so far that one must measure disease activity, one must measure patient-related quality of life. One needs to measure damage at the beginning and at the end, and one needs to measure drug toxicity, and whether there is a separate responder index or whether a responder index is derived from these components is something that needs to be clarified, but at least all of these things need to be measured.

DR. ABRAMSON: Dr. Harris.

DR. HARRIS: When Matt made his comment about starting organ by organ, I hadn't thought about it, but now I am, in fact, thinking with respect to what parameters should be measured.

I actually am more persuaded that this time perhaps we should go organ by organ, in which case one need not include, as I see it, all parameters. It depends on which organ you are looking at. I am very in favor now of saying, well, if we are looking at something that affects that integument, then, there are specific features relative to that, the kidney, and so on, because these global measurements, I feel, where everything is included, I feel uncomfortable about that.

DR. ABRAMSON: Did you mean that if you have a primary outcome like skin or kidney, that you would not

collect data for one of these other disease activities?

DR. HARRIS: I may not. Certainly, if I am looking at the skin alone, I can imagine that there are specific things that might deal with the skin in lupus, that may affect nothing else. I don't feel as if there should be an obligation to do so.

DR. ABRAMSON: Dr. Gladman.

DR. GLADMAN: I think for anybody who looks after lupus patients not to examine and record the whole patient is ludicrous. The disease activity indices, it doesn't matter which one you use because they all measure the same thing. There are only so many parts to the body, there are only so many systems that we have, and all of them get involved in the disease process.

So, the reason the instruments are all so comfortable is because they are all measuring the same things, they just give it a little bit of different weight, but the recording needs to occur for each patient. What you use as your outcome can be varied, and in individual patients, the outcomes may be different.

One patient may have renal disease, another patient may have CNS disease, and you may choose to look only at patients with individual organs, but I think that in terms of recording, the whole patient needs to be recorded.

DR. ABRAMSON: Dr. Fortin.

DR. FORTIN: I certainly cannot agree more, but I would like to also add if we are really trying to focus on which measures, I think I agree with Dafna at the beginning and at the end, at least to describe the patients. They may not show any differences in a short randomized trial, but disease activity could capture the notions of flares and responders index in some of those instruments.

I know Dr. Buyon made a point at the SLEDAI meeting that had changed, and that was quite well demonstrated on her slide, but the SLEDAI has fixed weights to each item, so you don't have a choice of the score, but some other disease activity instruments have a breadth in the way you can score each item, so you could actually measure response, and you could actually measure a flare.

DR. ABRAMSON: In terms of we will be answering this question for the members of the FDA, let me make a statement and see if people agree.

We would have the consensus that a damage index like SLICC would be included, that a quality of life index like the SF36 would be included, and we now have differences of opinion perhaps as to whether a SLEDAI or a BILAG or some other instrument would be included, but some disease activity index would be included.

We perhaps can discuss the pros and cons of those two instruments as part of this discussion, but first, Dr.

Silverman.

DR. SILVERMAN: I have a conceptual problem with not dividing it down by organ. I think Jack Klippel's study again, in lupus nephritis, showed us and we all know that lupus nephritis as an example isn't lupus nephritis. The outcome definitely differs from the outcome of DPGN.

How can we lump these things together? So, I think that an indication for skin disease, for arthritis, is a very valid indication in lupus. Lupus isn't a disease, and here we are trying to fit everything into a single disease.

Now RA -- it is easy for me to say because I am a pediatrician -- is a more homogeneous disease. So, I think the idea of going for specific indications is very, very valid, and measuring each patient as a responder or not is the way to go because if we go lumping, we may again throw out drugs that may have specific indications, hydroxychloroquine being an example, may or may not be effective, but if it is, it is in a specific group of patients.

DR. ABRAMSON: Dr. Klippel.

DR. KLIPPEL: I think anybody who has taken the microphone has made the point that this is a process that is in evolution, and the perfect instrument isn't there, and that is going to continue for who knows how many more months

or years.

One of the things that concerns me, when I was listening to Jill Buyon talk about the modified SLEDAI, let's just say you chose to collect information using the SLEDAI, I am not sure how you would get from a SLEDAI to a RIFLE when you were using her example, because you wouldn't have the data that the joints went from 4 to 2. I mean you would be collecting --

DR. BUYON: Marc collected that data. I really think the point is that --

DR. KLIPPEL: Let me finish, Jill. When people give me a SLEDAI, I am supposed to check a box, you know, that says 1, 2, or 3 or something, and I thought the point of your exercise was I would check a 3 or whatever the number I would write down, but the joints only went from 4 to 2, so I wasn't actually picking up improvement because I wrote down 3.

DR. BUYON: Actually, that was the point. So, I think what we are saying here is yes, we understand the concept of organ specificity, but I believe what we are all agreeing to is that that may be an outcome measure, that may be the primary outcome measure, but you need to look at all of these different things, otherwise you might miss something. Who could have predicted necessarily that Plaquenil would be good for what it is good for?

So, what we are suggesting is a combination of things to be looked at. So, you looked at the SLEDAI for one thing, that is, flares that occur over time during your study. The RIFLE says where did you start and where did you end, and did you win in a particular organ system.

DR. KLIPPEL: So, Jill, I am arguing for RIFLE, so stay calm. But the point is, is that all the instruments that I have seen, I have never filled one out that says how many joints are actively involved, is it 4, is it 2. It asks me is the rash active or not, and then I have to use some definition, and as you pointed out, I would be entering something that you -- unless I am now going to collect RIFLE -- I am entering data that can't go back and be constructed into a responder index because I am locked in. I have never filled out one of the RIFLE things.

DR. BUYON: Well, no one has filed out RIFLE yet, because that doesn't exist in a clinical trial. But also what we didn't mention is that whatever instrument you choose, you don't just fill in the box. If you don't have a physician encounter form that gives you a level of activity, you have done nothing, and that is why you need to be an educated investigator. No company at all, no monitor is going to accept a SLEDAI without a physician encounter form.

So, if you wanted to derive back to BILAG, you might very well be able to. That is the point.

DR. KLIPPEL: Can I ask the OMERACT people who have given us the core set, what form am I filling out?

DR. STRAND: Well, the agreement that we had suggested is that you pick whichever disease activity index you would like to use, and you use it, but in the computerized Bombardieri thing, you answer the questions, and then it calculates the disease activity index, each one of them.

So, you can then have it written down in the form, in the chart, or whatever else, you are going to follow this patient according to SLAM, but if we look back, we will have BILAG calculated for each encounter, or we will have SLEDAI calculated for each encounter.

I think the issue there is that disease activity indices are helpful, but they do not reflect everything that we are looking for as an outcome measurement.

DR. ABRAMSON: Dr. Katona.

DR. KATONA: I am sitting here and telling myself that thank God that we are living in the computer age and that this is going to be a job that you going to have to be doing in the 21st Century, because I just can't see any easy answer.

I am listening to everyone and would characterize there is probably one group which is more the organ-specific group, the other one is the general lupus group, and I think

everybody is right. I mean I have not heard one comment here which I did not agree with today.

It is hard to summarize and hard to get drugs for that patient, and I think it is not going to be one drug, it is not going to be five drugs, it is probably going to be 10 or 15 drugs, and it is probably going to be like a sushi bar what you are going to wind up with, and the only way we are going to know it, which one to choose, if today we design the instruments and our goals to keep it really wide, I think we have to give so much latitude to the companies, but require them to collect the data because that is the only way we are going to know it.

So, I think flexibility and inclusibility, as well as education of the investigators is going to be the key, and I just would like to make sure we don't make it too narrow today.

DR. ABRAMSON: Dr. White.

DR. WHITE: My comment is somewhat similar. I would think that in terms of the question, what should be measured in all clinical trials, I am not sure we can do that. I think it might depend what claim they are going on, and I would rather see can there be agreement, if you are going for this claim, what should be measured.

It would seem to me that if you are going for a claim of a potentially disease modifier global agent, then,

that is a different set and clearly ought to be required. If you are going for something a little more limited in its efficacy, you might encourage the sponsors to gather all this other data, but given the cost and time, it might be difficult for the FDA to require that those other data on global issues be accumulated for a very specific issue.

DR. SHERRER: I have a question that is going back to what Dr. Klippel said in terms of the practicalities of filling out some of our measures, and I had the same problem.

In going back to the question, for instance, of the joint count on the SLEDAI, if you miss that fall from 6 joints to 3 because you basically checked off a summary or an assessment, I don't see how you could recover that, at least in the studies that I do, because contrary to what we do on the RA patients, in the lupus studies we don't do which actual joint count, and you could recover that later on lupus patients if you did do a joint count, but if you are just checking off an assessment that doesn't have that sensitivity.

The other thing is I just wanted to say that as a practitioner, I think patients consider it very important that we look at global measures of quality of life even when or as we also look at specific, organ-specific outcomes, because patients are interested in their whole being, and

not exclusively interested in one aspect, although that is important to them and to us.

So, I really do think that every study should look at, at least one global measure.

DR. PUCINO: One other issue would be the use of Phase IV trials if we are talking only eight randomized, double-blind trials. We need some long term observations.

DR. ABRAMSON: Dr. Ginzler.

DR. GINZLER: I think it is very important to distinguish between taking care of the patient, which needs to occur even in the course of a clinical trial, that patient is still an individual who you are treating and filling out the appropriate documents.

As Jill pointed out everything starts from the individual patient and the source document, and I am sure that Ken Schwartz could speak to this from having gone through every source document for the DHEA trials in minute detail.

Yes, of course, the instruments, because they are computerized instruments, do strip away some of the individual information, but it has to be there, it has to be back in the original source document, so you don't fill these out in a vacuum. You take care of the patient and you document, just like you can't bill for things unless you document. Those things have to be in the source document.

DR. SHERRER: Absolutely, because they are not in the computer.

DR. GINZLER: Well, if they are not in the computer, it is because that particular instrument was not designed to include them. Then, what that says is that instrument may or may not be useful for every possible purpose.

But if the information is available, one could presumably go back to the medical record, to the source document, and fill out retrospectively a new instrument. I am not saying that that will always be there and that you will have always considered every aspect, but if you have a patient with active arthritis, hopefully, you don't just write down patient has active arthritis.

You write down which joints are involved, whether they are tender or swollen, red, et cetera, and you can go back to that information if you have a good source document, and that is a necessary part of every clinical trial.

DR. ABRAMSON: Dr. Petri.

DR. PETRI: I think people have become very unhappy because we have the lumpers versus the splitters, and I think in lupus clinical trials, we can have both. I want to give again the RIFLE as an example, because the philosophy of the RIFLE was it could be interpreted in two ways. Let's call one way the global way, does global lupus

activity get better.

That could be determined from the RIFLE simply by telling people what percent of patients improved in at least one organ system, and didn't worsen in any.

So, let's say for drug X that answer turns out to be 40 percent. That means if you have a patient with multisystem disease and you look at drug X, you can say to your patient you have about a 40 percent chance that this drug will help you.

But from the RIFLE, you can do the other thing, as well, the splitting, the organ-specific response. So, from the RIFLE, you can determine for drug X that 80 percent of people with cutaneous lupus got better, but only 5 percent of people with renal lupus. Again, that is very useful information for both the physician and the patient.

So, with instruments that we are already thinking about right now, you can get both, and both are important for both the physician and the patient.

DR. ABRAMSON: As we close this question, may I just ask perhaps Michelle and/or Dafna, on the one hand, and David Isenberg, on the other, to list two of the major advantages of each of these disease activity indices, the BILAG versus the SLEDAI, and two of its disadvantages.

David, do you want to start, what is the best part about BILAG and the weakest part about BILAG?

DR. ISENBERG: First, just to back up a little bit, just to reiterate, that we already have in existence the mechanism, a computerized mechanism whereby from one single form, which you fill in when you see a patient, you can derive all of the major global score indices and also, if you want it, the BILAG index.

The BILAG index offers you, as I said in my talk, a kind of at-a-glance view of what is going on in the patient both at the time that you see them in the clinic, and also looking back over the previous visits.

It enables you to make very simple comparisons with serologic data, if that is what you want to do, and although it wasn't really designed as a drug responder index, we are actually pretty much using it that way at the moment in this study that I told you about, of looking at azathioprine versus Neoral.

Disadvantages, well, if you are not used to it, it seems a little daunting to start with, I have one of the forms that we use here, but like most of these forms, once you have done it a dozen times, it really becomes pretty simple and very easy to use. We have been using it in the UK and in some centers overseas for some 10 years now.

So, we have a lot of data. It has been validated. It has been shown to have construct validity, and so forth. We think it could be adapted relatively easily for these

sorts of studies.

DR. ABRAMSON: Michelle.

DR. PETRI: Let me start out by saying that no one is at the point where we know what is the best disease activity instrument, and that some may be more appropriate in some situations than others, but I wanted to address David Felson's concern about transition versus state.

One thing that is an advantage of the SELENA SLEDAI is it truly is a state instrument. You can fill that out in a patient you had never seen before. The other great advantage I think of the SELENA SLEDAI is it is as objective as the human beings in the SELENA trial could make it.

The interclass correlation coefficient is very high. A big problem with the more complicated instruments is that we start disagreeing among ourselves on how to fill them out.

So, I don't want anyone to think the SELENA SLEDAI is perfect. There are a limited number of descriptors, it can miss some things. So, you know, you lose some, you lose some, you gain something, and in the correlation coefficient you may lose some in that it is not completely all inclusive. So, I think in a lot of the clinical trials we need to have some of these instruments go head to head, so we learn more about them.

DR. ABRAMSON: Thank you. I think I see a

consensus conference in Cancun coming up.

DR. SCHWIETERMAN: Dr. Abramson, I think this has been a very helpful discussion. I think that the issues are becoming clearly framed, and I would like to thank all the participants for that, but the one thing I would like perhaps some additional information on, if the agency were to consider organ-specific claims, aside from nephritis, which is an obvious one, which others ought to be considered more heavily than others, or are the data not yet there to make that call?

DR. PETRI: I would say skin, joints, and, you know, I think what the patients want is the constitutional/fatigue/fibromyalgia drug.

DR. ABRAMSON: We may get into that a little bit later in the afternoon in a discussion on organs.

DR. McCONNELL: Could I ask a question?

DR. ABRAMSON: Yes, sir.

DR. McCONNELL: If you use the single form that allows you to get all five activity instruments concurrently, how long does it take to fill it out?

DR. GLADMAN: It takes the same length of time that it takes to see the patient. I mean it takes about 20 to 30 minutes to assess a patient with lupus completely. This is provided you have known them.

The first time it may take an hour. But a follow-

up visit on a patient with lupus takes 20 to 30 minutes, and it takes 2 minutes to fill the questionnaire once you have actually assessed the patient.

The more you have done, the less time it takes to check the spots in, unless you do it while you are examining the patient, while you are questioning and examining the patient, at which time your 30 minutes is your 30 minutes.

DR. ABRAMSON: No. 5. What serological and laboratory markers are useful in the outcome assessment for SLE clinical trials? These are clearly important biological markers of disease, but what are useful and should be captured?

Dr. Isenberg.

DR. ISENBERG: Michelle has told us very clearly this morning about the concerns that exist about the classic ones, which are to say the complement C3/C4 and DS-DNA. She has also touched on the subject of the particular assay that is used, measuring DNA antibodies is clearly critical, and I personally would certainly agree with that.

I think of the complement components, it is probably true that the breakdown component C3D/C4D are probably better than just the basic C3/C4 although they are not as widely available.

What we have been doing using the BILAG in the current study is to measure double-stranded DNA and C3/C4

and C3D/C4D where they are available.

DR. ABRAMSON: Dr. Gladman.

DR. GLADMAN: For the calculations of the SLEDAI, one certainly requires complement components, either CH50, C3/C4, and DNA antibodies. So, if one were to use these instruments, those two serologic markers are certainly necessary. There may be others that it would be more important. The C4D and C3D didn't make it into the SLEDAI because they weren't available.

DR. ABRAMSON: My own view is that while we don't know yet how useful they are in terms of predicting disease, and that is under study, to not capture this information is like treating a diabetic without knowing what the blood sugar is with respect to what that means to the disease pathogenesis.

Hopefully, we have some studies going on over the last year that maybe in a year or two we will have some data piggybacked onto the SELENA to look at these various analytes, but I am not sure what is useful right now.

DR. FELSON: I wanted to answer that question with a single word - none, strongly stated. I think there is evidence presented today and through many other studies that it is not appropriate to include as any measure of SLE disease activity any of the measures mentioned.

They don't correlate highly with what matters,

which is activity in particular organs or clinical manifestations in order to include them as measures in clinical trials.

I think you are commenting and I think everyone else is on their promise as potential surrogate measures when perhaps better identified and better studied, and I don't think anyone would disagree with that, but these are not gold standard measures of what matters in SLE. These are measures we use to clinically anticipate someone's potential flare.

DR. ABRAMSON: I think Michelle commented on this also. One of the problems with the literature is that we don't have -- when you do a randomized clinical trial, people are being seen every month or so.

The literature is filled with complement determinations, three-month basis, when they come to clinic, and it is very hard to put together that literature and make predictive value validity out of those data, but that needs to get done.

David.

DR. ISENBERG: I wouldn't entirely agree with what David said, although I think he is right, and I think those of us who have formulated the indices have been very careful to make sure that DNA antibody levels and C3 levels score either nothing or very little. We are testing against

those. But there are quite a number of studies in the literature which have suggested that there are correlations, for example, between DNA antibody levels and nephritis measured serially.

Certainly, the patients that I worry about, the patients whose C3 is serially falling, and whose DNA binding is serially rising, and although it is a little controversial, most of you will be aware of the studies that were published in Lancet by Bootsma and his colleagues some years ago suggesting that those patients whose DNA antibody levels are raising rapidly over a period of several months are the ones who are most likely to flare.

DR. ABRAMSON: I think it is the deltas that are very important that haven't been captured, as well.

DR. KALUNIAN: It is also important to look at those surrogate markers if that is what the mechanism of your drug's action is. I mean if you have a B-cell tolerogen that its role is to downregulate the DNA response, then, obviously, it becomes that much more important to look at that.

DR. ABRAMSON: Dr. Fernandez-Madrid.

DR. FERNANDEZ-MADRID: I think that all the studies that look at the flare, a renal flare, exclude patients that don't have anti-native DNA antibodies, and we have followed a number of patients, quite large, with

systemic lupus erythematosus without anti-native DNA antibodies, and I think there is a lot that we don't know about other antigen antibody systems in lupus nephritis.

DR. ABRAMSON: Are there any other comments?

If not, we will break until 1:15. Thank you very much.

[Whereupon, at 12:15 p.m., a luncheon recess was taken.]

A F T E R N O O N P R O C E E D I N G S

[1:25 p.m.]

DR. ABRAMSON: I would like to begin the program and turn the floor over to Dr. Silverman who had a comment that he would like to make following this morning's session.

DR. SILVERMAN: I guess the major comment I want to make is for the inclusion, at least not initially, but certainly to include it in late phase III or early phase IV trials is the inclusion of pediatric patients. They do actually make up 20 percent but, more importantly, from a toxicity point of view, you also have a population where the other confounding factors generally are not there such as when we were looking at the role of antiphospholipid antibiotics and deep-vein thrombosis, our patients don't have an added risk.

In addition, we also have the population who is going to have the disease the longest so I think both from a scientific as well as a health-outcome point of view, I just want to make the plea to include not early but certainly into every trial built in to phase II or phase IV the inclusion of pediatric patients and also the inclusion of patients both prepubertal and postpubertal.

DR. ABRAMSON: Thank you. We will begin the afternoon which is focussing on clinical trials in lupus nephritis and other organ-specific manifestations of SLE.

Our first speaker will be Dr. Susan Ellenberg on surrogate markers.

Surrogate markers

DR. ELLENBERG: Good afternoon.

[Slide.]

First, a disclaimer. I have not worked in this area of disease so I am not going to talk about specific issues in lupus. This is a general presentation on considerations in the use of surrogate endpoints.

[Slide.]

This is a clinical definition of surrogate endpoint. I think it is useful. "A laboratory measurement or a physical sign used as a substitute for a clinically meaningful endpoint that measures directly how a patient feels, functions or survives." The implication is that the surrogate does not measure those things. It is something that really doesn't have anything to do with how a patient is doing at that moment in time in terms of how the patient feels.

[Slide.]

This is a more statistical definition. This was put forward by Ross Prentice in 1989. "A surrogate endpoint is something that is known to be prognostic for the clinical outcome of interest." And this is the key, that, "The treatment effect on the surrogate provides full information

about the treatment effect on clinical outcome." That is all the information about the effective treatment is contained in the surrogate.

This is a very demanding criterion and one that I am not sure anybody thinks will ever be met by any possible surrogate. But conceptually, I think it is useful in terms of what we would really like to think of as a surrogate endpoint.

[Slide.]

We all know why we are interested in using surrogate endpoints. The results can be made available faster. We can identify good therapies earlier. Your study is cleaner when you do a shorter study. You have less non-compliance, less dropout. An issue particularly when you are looking for a surrogate for survival is that the patients in the trial, themselves, are more likely to benefit if you can identify a treatment that prevents death and you can identify it before there are any deaths observed in the trial, then even the people in the trial who are, perhaps, getting a placebo or a less effective therapy are going to benefit.

Faster studies are cheaper. Sometimes, you look for a surrogate when the true endpoint is something that is evasive, difficult to measure and it is less subject to influence of extraneous factors like changes in standard of

care or competing risks.

[Slide.]

In drug development, surrogates have essentially always been used to identify new agents with promising biological activity. In fact, they are developed specifically to have particular biological activity so we are often not surprised when they do.

They can be used for prioritizing active agents for a definitive study. If there are a number of things that are going through phase I, the ones that seem to have the greatest activity and the strongest effect on the biologically relevant endpoints are going to be the ones that go into further studies faster.

The areas where there has been more controversy have been in actually assessing efficacy of new drugs and supporting drug approvals and then, ultimately, comparing active or effective treatments. Those are more difficult.

[Slide.]

A lot of markers have been used, have been considered surrogate endpoints. Some of these are arguable. Some people would say, "That is not a surrogate endpoint. That is the real endpoint." Things like tumor size and viral load, people have made those arguments. But these are all parameters that, in terms of the patient at any moment in time, wouldn't really affect the way that patient was

feeling except, perhaps, psychologically, if they knew what the value of that endpoint was.

[Slide.]

The fundamental problem of surrogate endpoints and the thing that gives us pause is that we recognize they may be associated with a clinical outcome but not causally; that is, by affecting the surrogate, you may not, ultimately, affect the clinical endpoint.

Furthermore, when you measure the surrogate, you don't account for adverse effects which may be going through some other pathway which may cancel out part or all of the apparent treatment benefit or whether it is making the actual endpoint that you want to measure, the efficacy endpoint, worse or whether it is making something else worse that is counterbalancing the benefit you get from what you are trying to do.

Another problem, and we have seen this from time to time, is that a surrogate doesn't account for beneficial effects of a drug which might occur by a pathway that doesn't include the surrogate. There have been examples where we would have been misled if we hadn't gotten the clinical endpoint because there was no effect on the surrogate but there was a strong effect on the clinical outcome. So we worry about that a little bit, too.

[Slide.]

This is the big question; how do we validate surrogate endpoints? Identifying is not so hard but validating is hard. These are sort of things that we intuitively consider, obviously a biologic rationale based on the known mechanisms of the drug and the known pathways of disease, if we have good natural-history data, we can assess--we know which markers have some prognostic value.

If there have been a series of studies, if there has been a history of drug development, one would want to know whether these markers respond to effective drugs. Ultimately, you need to know that there will be a correlation of treatment-induced effects on the surrogate with the effects on the true endpoint.

If you improve the surrogate, do you see improvement on the clinical outcome in a variety of circumstances?

[Slide.]

Those are all sort of natural and intuitive. There has been a lot of effort over the last ten or twelve years in thinking about how to statistically validate surrogate endpoints.

Prentice's criterion which is that you have got to show that all information on the clinical effect is mediated through the surrogate is, as I said, very restrictive. In 1992, Freedman and his colleagues at the National Cancer

Institute proposed that, "Well, maybe we don't need to show that all the information is there but we can estimate the proportion of information on the clinical effect that is mediated through the surrogate and if that is high enough, then maybe we can consider that as a surrogate."

There was a workshop at NIH last December on statistical methods for evaluating surrogate endpoints. There were a number of people who talked about modeling approaches, basically, rather than considering the question of validation, just estimating the treatment effect on the true endpoint from the data on the surrogate.

We heard of a variety of presentations in different disease areas. These were mostly coming from studies where data were collected on both types of endpoints, surrogate endpoints as well as the clinical outcomes.

In some cases, there was very little clinical data and people used metaanalytic approaches to combine data from studies that had been done to model the relationship between the proposed surrogate and the clinical outcome. These got pretty complicated so that it got to be sort of a black-box kind of thing. It was not intuitive what was actually going on.

Others at this workshop proposed doing some direct modeling of what was happened, pathophysiological modeling,

where you tried to actually model what was happening, the dynamics of the treatment on the disease process. That work is much more preliminary, but there is a lot of excitement about that.

[Slide.]

One of the vexing problems in all of this is knowing, even if we define a surrogate with a particular drug, how could we be sure that it was going to work for another drug, whether it was in the same class but a little but different or even in a different class.

There may be different causal pathways for efficacy. There may be different toxic effects. So even if we get to the point where we think we have a valid surrogate based on the data that we have, we are always going to be a little bit uncomfortable about whether it is going to apply to the next new treatment that comes along.

[Slide.]

So, what do the regulations say about surrogate endpoints. Interestingly enough, when I was putting some material together for an earlier talk, I discovered this section in the Biologics regulations. I don't quite know how far back it goes, but it certainly predated the accelerated approval regulations.

It says that, "Effectiveness means product must serve a clinically significant function, must be

demonstrated in controlled clinical investigations, but serological response data may be used to demonstrate effectiveness when a previously accepted correlation between data generated in this way and clinical effectiveness already exist."

This sounds like all the surrogate endpoint stuff that we have been talking about. This is in there almost surely because of the vaccine issue and the fact that vaccines are often evaluated based on serum antibody levels. But I guess we were ahead of the curve here.

[Slide.]

In 1991, the accelerated approval regulation was proposed by FDA and this was made final in 1992. It is limited to serious and life-threatening diseases and it provided for the marketing of drugs shown to have a positive effect on a surrogate endpoint where there is real potential for advantage over existing therapeutic options. This was never meant to be for "me too" kinds of drugs.

There is a requirement that studies to evaluate the clinical effective treatment be ongoing; that is, there is an expectation that ultimately the effect of a treatment on the clinical parameters of real interest will be completed successfully and it also provided the drugs ultimately found to have no clinical effect could be withdrawn.

[Slide.]

Why did this come up in 1991 and 1992? At that point, or I guess for decades before, a relationship between some of these surrogate measures and clinical outcomes was implicit for a large number of products. Nobody worried too much about blood pressure as a surrogate endpoint. Lots of drugs were approved for controlling blood pressure.

Drugs had been approved for shrinking tumors. There wasn't a lot of angst about surrogate endpoints until a few dramatic counter-examples arose in the late 1980s and possibly earlier. An example that most of you are probably familiar with is the antiarrhythmic example where a drug that prevented arrhythmias in post-myocardial-infarction patients was found to not only not improve survival, as had been expected, but drastically worsened mortality.

That was sort of a shakeup. I think it made people start to think more surrogate endpoints. There were other examples in the cardiovascular field that produced counterintuitive results in terms of mortality. So there was a lot of rethinking about how much we should be relying on surrogate endpoints.

Almost at the same time, the emergence of AIDS created intense pressures to make promising drugs rapidly available. So you have these two opposing things, a realization that surrogates were problematic together with

an incredible urgency to get AIDS drugs out sooner.

[Slide.]

What has been the history? I have been able to count twenty-seven--I don't guarantee that is not plus or minus something, but that is what I have been able to come up with--twenty-seven accelerated approvals through the end of 1998. These have been mostly AIDS and cancer drugs but there have been some important exceptions.

I don't believe that any of the drugs that have been approved by accelerated approval have actually been withdrawn from the market due to failure to confirm the results with clinical studies.

[Slide.]

I just have a couple more comments on regulatory documents. The International Conference of Harmonization produced a document entitled Statistical Principles for Clinical Trials. This was published a few months ago as an FDA guidance document. It discusses, among other things, the issue of surrogate endpoints in clinical trials.

It states that, "This may be the primary basis of evaluation in some cases." It notes that caution is needed. It reminds us that it won't always predict accurately clinical outcome and acknowledges that we don't have--we are not ready to set forward a statistical criterion for validating a surrogate and that evaluation case-by-case is

unavoidable at the present time.

[Slide.]

Finally, surrogate measures are explicitly measured in the guidance document for rheumatoid arthritis that was just released. This document specifies radiographic progression of disease as a potential surrogate endpoint but notes that phase IV studies demonstrating ultimate effects on clinical endpoints will still be necessary.

I will stop there. Thank you.

DR. ABRAMSON: Thank you very much.

Are there any questions for Dr. Ellenberg?

The next speaker will be Dimitrios Boumpas from the National Institute of Arthritis and Musculoskeletal and Skin Disease on lupus nephritis.

Lupus Nephritis

DR. BOUMPAS: Good afternoon. I would like to thank the organizers of this meeting, Bill Swieterman and Jeffery Siegel, for inviting me over. Jack Klippel was one of the most practical people this morning. I will try to follow his example. I will give you specific answers and, I hope, clear answers.

I am not a nephrologist. I am a rheumatologist. But Jim Balow has made me an honorary nephrologist. It is a good thing to see in this panel both nephrologists and

rheumatologists. I think it is very healthy. It is an interaction that we have enjoyed over the years at NIH which has been very fruitful.

[Slide.]

This is a list of outcome measures that somebody could use for lupus nephritis. You will see here most of the things that were discussed in the morning session. I guess everybody has agreed that outcome measures for lupus nephritis have to be scientifically valid, they have to be clinically relevant and they have to be, also, feasible and practical.

It is very difficult to find criteria that meet all of these three requirements. The previous speakers this morning have discussed things like lupus disease activity, lupus activity measures, or things like the damage index or the changes in the health-related quality level.

These are things that I know the list about so I will not comment on them. But I think, based on the previous discussion this morning, I would feel comfortable if they were included as secondary tertiary outcome measures in trials of lupus nephritis.

[Slide.]

Another general comment before I start discussing outcome measures in lupus nephritis. Although I understand that this discussion is focused on large, definitive

studies, I think it is useful and practical to distinguish the lupus trials in the small pilot studies which will allow you to investigate the potential usefulness of a new, exciting biological agent and, also, the large, more definitive studies.

I will try to give you some of the inclusion criteria that we have tried to develop over the last year. Also, I would like to start this discussion by saying that the natural history of membranous lupus nephritis and proliferative lupus nephritis is different, I don't think it is a good idea lumping them together unless there is a way of stratifying the criteria and unless you consider, in your outcome criteria, criteria that will be meaningful to the membranous versus the proliferative because I am not sure that you could use, necessarily, the same outcome criteria for these two types of lupus nephritis.

Finally, and that is something that I hope we have some time to discuss at the end of this, I think one of the problems that we have with lupus is how to distinguish active, ongoing neurological injury to the kidney from fixed, irreversible injury. I am referring by that to the fixed proteinuria and the fixed hematuria and how somebody could sort them.

It is very useful to sort them out because otherwise you compromise the validity of the study.

I will not discuss inclusion criteria for lupus nephritis. In the supplement to lupus or lupus nephritis, Jim Bell and I with the help of the rest of the team at NIH tried to discuss some of the outcome measures in lupus nephritis and some of the inclusion criteria. It was published in December. I don't have time to go over that but, for those of you who may be interested, you can look for more details there.

I promised to you that I will be specific and clear. It was very helpful that Jeff gave us clear questions. I will just answer your questions.

[Slide.]

These are the questions that I was asked to discuss; indicators of clinical benefit and to discuss the role of the doubling of the serum creatinine, normalization of the elevated serum creatinine, the decrease in the frequency of renal flares and decrease in proteinuria. Also I will discuss the decrease in the use of toxic agents as an outcome and also the issue of acceptable control regimens.

[Slide.]

Starting with the doubling of the serum creatinine, this is something that we have traditionally used at NIH. It reflects at least a 50 percent fall in the GFR. It is not clear whether somebody could use a smaller increase in the serum creatinine, that is 30 percent or

50 percent.

We tried to answer that retrospectively in our lupus cohort but we did not have enough data. We just finished an extended follow up of the patients who participated in the last two trials at NIH and, hopefully, we may be able to publish something on the validity of the 50 percent or the 30 percent change in the serum creatinine.

We have maintained that absolute thresholds are less reliable since they reflect variable change in the GFR. Some people, and I am sure we will have some discussion, have used a cutoff, for example, of patients who reach a serum creatinine of 2 or 3 as an outcome.

We don't like that because if you start, before the study--if you set the threshold at 2, and you start with a creatinine of 1.7 and another patient starts with a creatinine of 0.8, I am not sure that you are comparing the same thing. So, for that reason, we have stayed away from this and we like the definition that the change in renal function is doubling of the serum creatinine.

Also, Jim Balow here has used, and other people have used, the annual decline in GFR or the slope of 1 versus creatinine, or the trial in IGA nephropathy for insulin-dependent diabetes mellitus, but the natural history of the disease in lupus is different and I am not sure that they are as reliable.

[Slide.]

This is a study that Jim Balow presented one or two years ago in a similar FDA meeting. This is comparing the doubling of the serum creatinine versus the end-stage renal disease. The initial NIH studies used end-stage renal disease as an outcome criterion. As you can see here, in order for 20 percent of the patients to reach this outcome, you should wait for at least eight years.

You see that doubling of the serum creatinine pretty much parallels the course of the end-stage renal disease but the advantage of that is that you can decrease the follow up to achieve this--for 20 percent of the patients who achieve this outcome from eight years to four years.

I will predict that if you are using a 50 percent or 60 percent increase in serum creatinine, you may get a similar slope. We are, as I told you, are just in the process of generating the data accumulated by completing the longer follow up of our patients.

So I think, as a single outcome criteria, doubling of the serum creatinine is certainly acceptable and it is clinically meaningful. How practical it is and how feasible is something that we can discuss but I am not sure that there is anything better right now.

[Slide.]

The next issue that I would like to discuss is the issue of the normalization of the serum creatinine. Ed Lewis is going to, hopefully, discuss their experience. They reported in their lupus nephritis collaborative study in their Annals paper that the normalization of the serum creatinine within 48 weeks predicts a lower risk for renal failure.

I think it is a useful outcome criterion but has limitations. Therefore, I don't like it as a primary outcome criteria. Probably we could consider it as a secondary or a tertiary for the following two reasons. The first one is that patients may start with a different--that may reflect different changes in the serum creatinine.

For example, you may have a patient who took prednisone who starts from a creatinine of 2.0 and then normalizes the creatinine to 1.2 and the patient who takes an experimental drug who starts with a creatinine of 1.5 and decreases to 1.2. I don't think that the efficacy of the two drugs is similar but this will count as similar. So that is a limitation.

The other limitation is, in contrast to the lupus collaborative study, the prevalence of the abnormal serum creatinine. I am talking about the proliferative lupus nephritis. In the last proliferative lupus nephritis trial at NIH that Mark Gurley published in the Annals in 1996,

approximately one-third of the patients had an abnormal creatinine at baseline and this is in contrast to the 60 or 70 percent in Dr. Lewis' collaborative study.

So I am not sure, because of the prevalence and the other problem that I discussed--I don't like it as much and that is why I think it would be important to report it but, again, as a secondary or tertiary outcome criterion.

[Slide.]

The issue of the normalization and the normal serum creatinine, I think is an important one not only because of the normalization of serum creatinine six months into the study predicts a good outcome but also having a normal serum creatinine, that, in the long term, is also predictive of a good outcome.

This is an analysis that Barry Fessler with Howard did in our lupus nephritis trial that was published in 1992. As you can see here, if the patient has normal creatinine at six months into the study--and here we have two groups, one group who took IV metal for six months versus a group which took IV Cytoxan for six months.

The treatment was stopped. Again, having normal serum creatinine at the end of the six months, at the end of the treatment, that is a good prognosis.

[Slide.]

The decrease in proteinuria; this is the most

difficult of the outcome measures in the context of the proliferative lupus nephritis. For the membranous lupus nephritis, it is easier. I am glad there are people in the audience like Howard Dorstin who can speak--who have experience with that. But, in the proliferative nephritis, I am not sure how to classify proteinuria.

It is very difficult to make a convincing case that proteinuria, by itself, will be a primary outcome in proliferative lupus nephritis, probably with one exception. I hope we have enough time to discuss that in the discussion period.

If you have a patient who has a diffuse proliferative lupus nephritis and has nephrotic-range proteinuria, if that patient decreases the proteinuria to below 2.0 grams, and that is without increasing the serum creatinine, in that case, I would be willing to discuss it as a primary outcome criteria.

Otherwise, I see the decrease in the proteinuria is a secondary outcome criteria or something that could be used as a composite in a responder index. But I will get back to that. So these are some general statements that proteinuria reflects the involvement of glomerular capillary groups, that nephrotic-range proteinuria is a mark of disease activity indication of therapy.

The problem is what do you do when you have lesser

amounts of proteinuria.

[Slide.]

Another problem that I see with proteinuria is that, at least in our experience, proteinuria rarely recedes to normal levels. We are using the term "fixed proteinuria." Probably that reflects the biases that NIH has. We see patients with more advance disease. They come to us after they have had other treatments. But, in most of our patients, the proteinuria does not go away and I don't think that is because you do not quiet down the disease.

We think that probably this is, in most cases, because we have fixed irreversible damage in the basement membrane.

This is a capricious test and there is a lot of day-to-day variability. You know that the proteinuria is not very practical, to do the 25-unit collections, especially to do three of them. The patients hate it and the investigators are not very happy about that.

[Slide.]

Another problem is proteinuria is how to define what is a clinically important change in the protein. There are many people here who propose different definitions. I like Moroni's, who works with Claudio Ponticelli in Milan, who has used in a paper a few years ago this definition of a significant change in proteinuria; "If the baseline is

nephrotic, at least 50 percent decrease; if the baseline is non-nephrotic, more than 2 grams per day."

We have seen, and other people have reported, that you have a reduction of proteinuria to less to 1 gram, that is always provided that the creatinine does not increase. That identifies a patient with favorable long-term prognosis. So the 1-gram threshold may be a clinically meaningful threshold.

Another thing about the proteinuria, and this gives relevance for small pilot studies, is that up to 50 or 60 or 70 percent, in some cases, of our patients, they have 50 percent reduction in proteinuria in six months. So if you are anxious to see a quick change in a pilot study, that is something that you could use in such pilot studies.

[Slide.]

This is just, again, from the same analysis from the patients with severe proliferative lupus nephritis treated with IV Cytoxan or IV metal for six months. You can see the resolution of proteinuria to less than 1 gram is associated with a good prognosis provided that the patient does not increase the serum creatinine.

[Slide.]

Decrease in the frequency of flares; the issue of flares and the issue of remission is very emotional. The rheumatologists, especially the lupologists, have strong

feelings about definitions. But I will try to be provocative just to stimulate some discussion.

It is a problem how to define a flare because pretty much you have to define what is a remission. The issue of the outcome is an important one, and the frequency. So let me just say a few words about these things before I discuss the usefulness of the flare.

[Slide.]

I like Ponticelli. Before I start, several people have used definitions of flares and remissions over the years. But I will concentrate on the ones that may be helpful for today's discussion. So this is what Claudio Ponticelli--I got this from something that he wrote in the Supplement in Lupus in 1998.

This is based on their experience with 70 patients who had a mixed bag of proliferative nephritis and membranous nephritis. What is different about Ponticelli's cohort with our cohort is that we do not have membranous patients in the studies that I will be describing to you.

30 percent of their patients were membranous. That is important to keep in mind. So 46 percent of their patients experienced more than one flare. These are patients who are treated usually with high-dose steroids for two or three months or with daily oral Cytoxan for three months. Then they are on maintenance steroids.

They treated all the flares with immunosuppression. They detected 25 proteinuric flares. By proteinuric flare, what they mean is you have at least a two-gram increase in the amount of protein excretion. If your baseline proteinuria is below, non-nephrotic, or a 50 percent increase in the proteinuria if your baseline is at the nephrotic range.

All patients were treated with immunosuppressive therapy. None of the patients with proteinuric flare developed--after a follow up of approximately ten years, all of them maintained stable renal function. They had 21 patients with nephritic flare. By that, they mean active urine sediment with mature serocast and also at least 30 percent reproducible increase in the serum creatinine.

From the 21 patients with the nephritic flare, ten doubled their serum creatinine. So the message from that was that the renal flare associated with impairing of function has certainly an important effect on the outcome.

As far as the incidence of the flares, this is their estimate. In their cohort with this particular treatment, they have 0.31 flares per patient per year and that includes both proteinuric and nephritic flares.

[Slide.]

We like Ponticelli's definition but we felt that we should try to improve it because that is what we like to

do at NIH. So we analyzed our experience with 74 patients. This is out of 120 patients who were treated with immunosuppressive therapy. 74 patients of these patients achieved remission defined as less than 1 gram of proteinuria with inactive urine sediment and no increase in the serum creatinine.

So we are able to document--Gabrielli able to document 31 flares in our cohort. These are patients with either pulse Cytoxan or pulse metal or combination. The medium time to the flare was 21 months.

46 percent of the patients who received methylprednisolone flared. Less was 20 percent after cyclophosphamide. We only had three proteinuric flares but this is because we do not have any membranous patients in this cohort. We divided the nephritic flares into three different categories; mild, just a reappearance of serocast or worsening urine sediment; moderate, when you have worsening of urine sediment but reappearance of serocast with hematuria plus at least 2 grams increase in a 25-unit protein excretion and severe--the ones that they have what I describe for moderate plus 30 percent at least increase in the serum creatinine.

We feel that this is an important distinction to make because clinically, if you see a patient who has the reappearance of serocast or a worsening urine sediment and

that is associated with increased proteinuria, I think you will feel differently this patient from the patient who just has the reappearance of serocast with no significant increase in proteinuria.

That is why we thought that would be a useful distinction to make. Two out of the three patients with severe flare developed end-state renal disease but, again, all patients, especially the patients with moderate and severe flare were treated aggressively with a high dose of steroids and, in most cases, with cyclophosphamide. We are just extending the follow up.

[Slide.]

So, in summary, just to give you a specific and clear answer, the flares are useful especially if they are moderate for proliferative lupus nephritis. But I don't think the mild flares, the way we define them, probably should be included. But the moderate to severe flares, I think they are important. They can be useful outcome measures for lupus nephritis but they are less common and occur late after intensive immunosuppressive therapy.

That is, if you give a prolonged course of immunosuppressive therapy like our NIH protocols are, very few patients will have them and so you will need a very extended study and many years of follow up and a large patient population to be able to give a satisfactory answer.

However, if you have--and I am aware of one European study that tries to address this, that after a fresh lupus nephritis, after a short course of Cytoxan for three months and then these patients are randomized to receive two different regimens. I will measure this in a meeting and I will not mention the drugs.

In such a design of the trial, you may be able to experience enough flares to be able to compare between two different drugs. But, still, if you do the sample calculations, you need at least 200 patients to carry such a study. But I like the flares, especially the nephritic and moderate to severe. I am not sure proteinuric flares for proliferative lupus nephritis is a good outcome, at least a primary outcome.

[Slide.]

Decrease in the use of cytotoxic agents. There is an easy and a difficult answer in this question. The easy answer is when things are clear-cut. For example, this is from our experience with IV Cytoxan. When you compare less than seven pulses or a short Cytoxan course versus a long course of Cytoxan, you can see the irreversible amenorrhea, 12 percent in the short Cytoxan versus 39 percent in the long Cytoxan. This is statistically significant.

That is the only statistics you are going to hear from me today. So, in this case, I think that it is very

clear if you have the regimen that was decreasing by 50 percent the pulse of Cytoxan and this is associated with a tangible benefit, it is clear that you have got something good.

The question is where you should set the threshold. How do you go from ten pulses to seven pulses. That is the different part of that. And that applies to the prednisone and I think the morning session illustrated some of the difficulties of some of these. So this is clinically meaningful.

I heard several clinicians say that that is important. I know that is important but it is difficult to standardize it and you don't want to use that as your primary outcome criteria. But it is problematic especially when the clinician is not blinded to the experimental therapy and this is the case for most lupus nephritis trials; so secondary, tertiary, or but not the primary outcome.

[Slide.]

The final issue is the control group. I think you have to distinguish what kind of trial you are trying to do. I understand that the oncologists are getting away from this concept of induction versus maintenance therapy but, at least for practical purposes for us, at least in lupus, this worked. So I still keep it, although I may be

anachronistic.

So, for induction therapy in a patient with fresh proliferative nephritis, I think you could use--you have one of the three choices; either high-dose steroids or pulse steroids, as some people do with moderate-dose steroids; daily oral Cytoxan for three months or a bit longer. That is a very popular regimen, at least in Europe; or a short course of Cytoxan, six passes of cyclophosphamide.

This will be like a control group. You may argue that cyclophosphamide is not FDA-approved. That is not my problem. But I think you could choose one of the three ones. If you don't like the cyclophosphamide, you could use steroids.

As far as the maintenance, you could use low-dose steroids. That is an acceptable dose for a rheumatologist, I would like to believe and also I hope for the nephrologist. But I know it is not for the hematologists. They like a higher dose of steroids for prolonged periods of time.

So this is what I would be advisable; low-dose steroids with this dose here, azathioprine or quarterly cyclophosphamide.

[Slide.]

This is something that I have the least understanding but it is a nice term. It is very fashionable

and it has worked for the lupus nephritis. I think initially I was a little bit skeptical about this but I think it is part of aging is just to be a little bit more open minded. So that is something that I think is worth considering. This is the use of the composite versus single-outcome criteria.

We have heard about the responder index which integrates several measures of outcome. People in the lupus trials have used this terms. They did not wait for the rheumatoid arthritis investigators to find these terms or these outcome measures.

There are people who have used improvement, response, remission, several investigators including Mark Gurley from our own group. So these are terms that have been used in the literature. The theoretical advantage if you use composite versus single-outcome criteria is that you may decrease the sample size and you may reach this outcome earlier which is highly desirable.

The problem that I have is that these are not validated. John Davis has started the process of trying to validate some of these composite outcome criteria but the outcome that we had when we started doing this analysis was not long enough and we just were completing the extension.

We may be able, retrospectively, to have some data. So it is an attractive idea. It is worth considering

but I am not sure--and I will give you an example of how we will apply it in the context of lupus nephritis or how we have applied it.

[Slide.]

The responder index from the outcome criteria, these are some of the things that somebody could consider in that.

[Slide.]

A few things about the urinalysis because this has been a very, very emotional--I have been trained at NIH and spending your life at NIH you think that doing an urinalysis, a good urinalysis, is pretty much easy. But when we try to collaborate with some investigators for multicenter studies, we realize that, in the outside worlds, the urinalysis is very problematic. It is not easy to do good urinalysis.

So in the center or community-based laboratories this is problematic, even in the context of multicenter trials, at least our experience has been that it is also very problematic and although I like the urinalysis and I was trained to look at the urine of the patients myself and just make the decision based on what the urine shows, I am here to say that urinalysis in the context of multicenter studies is not something easy. It is not practical. It is useful but it is not practical.

But I like the notion of using hematuria, hematuria with dysmorphic red blood cells as a categorical measure, either presence or absence. I think in that context it is useful as an inclusion criteria for patients with nephritis or part of an outcome measure, of a composite of outcome measures.

But determinations, semi-quantitative determinations, the urinalysis can be very messy if the red blood cells--even the red blood cells can be messy in multicenter studies for the serocast--people have done studies in pretty much the community-based laboratories, they miss them most of the time. So they don't work.

DR. ABRAMSON: Dr. Boumpas, can you come to a close soon, please.

DR. BOUMPAS: Yes. I'm finished.

These are the last two slides.

[Slide.]

The proposed core criteria for response of lupus nephritis. For small pilot studies, small changes in renal function, urine sediment, proteinuria and lupus glomerulonephritis, these are some of the things that you will consider.

[Slide.]

But for the large definitive studies, you have to use harder outcome criteria. If there are any questions, I

will show you how we have dealt with it and how we have defined remission and relapse.

I'm finished.

DR. ABRAMSON: Thank you very much.

I think we will hold questions because many of these same issues will come up when we address the questions put to us. So we will go to our next speaker, Dr. Edmund Lewis from Rush Presbyterian, St. Luke's Medical Center, also to speak on lupus nephritis.

Lupus Nephritis

DR. LEWIS: Thank you Dr. Siegel, in particular, for inviting me. I think this is a very important meeting and that is why I am here.

Whenever I am in a room with a bunch of rheumatologists, I have this feeling that I have to introduce myself. It is like an American Express ad. "You don't know me, but I'm the reason you are not still treating your lupus nephritis patients with plasmapheresis." How's that?

[Slide.]

I think Stewart put it well. Being a nephrologist and being preoccupied with other renal diseases over the last several years, I can say that, in terms of lupus nephritis, I believe the situation is absolutely appalling. We have not a lot of information about the maximum benefit

that patients could derive from therapy or even what the best therapy should be applied.

Despite a good deal of literature on the subject, I don't think that there are many people who are in total agreement about even whether alkylating agents should be added to high-dose prednisone in severe lupus nephritis.

The reason I am here is because I think that it is important that we don't keep multiplying this by having patients treated off-label with every immunosuppressive that comes down the pike or we will never know how to treat these patients.

[Slide.]

What I am going to do is review twenty-five years--well, ten years experience now with a group of patients that we entered into the plasmapheresis trial. They were 86 patients who had very well-documented severe lupus nephritis. We had a very rigid histologic criteria, a panel of four well-known pathologists and so forth.

That study ended in 1986. So we now have an average follow up of ten years. Mo Reichlin, last week, as a matter of fact, wrote me a letter accusing me of starting a new field of the archeology of lupus nephritis.

We have looked at the follow up of these patients and I think that there is some information that comes out of that that could be helpful in terms of the way one might

design a study or at least an approach to the way one might design a study.

Remember, all of these patients were treated with high-dose steroids for at least four weeks and oral Cytoxan for four weeks. And then, if they did well, the prednisone was tapered. Half of the patients were treated with plasmapheresis. There was no difference between the groups and so all of our data are pooled. This is the total of the 86 patients who came in.

As I say, it is very important when we talk about lupus nephritis to define that lesion and make sure the patients coming into the study have what you think is severe lupus nephritis. Otherwise, you get into the problem of the Will Rogers phenomenon, Will Rogers talking about the Okies when they left the Dust Belt to travel to California. He said, "When the Okies left Oklahoma and went to California, it raised the IQ in both states."

What that essentially means is that if you put patients with lesser renal disease into a several renal-disease study you will not get the answer because they will all respond to therapy.

[Slide.]

These patients were typical patients with severe lupus nephritis. They were 32 years old. They had, on the average, a serum creatinine of 1.9. The median was 1.5 and

they had 6 grams of proteinuria.

[Slide.]

At follow up, the average being 141 months of follow up at the time that this slide was made, 26 percent of these patients had gone on dialysis. Seventeen had died associated with their renal disease. 10 percent had non-renal deaths and 47 percent had stable renal function.

[Slide.]

The patients' survival, then, over the ten-year follow up, the blue line, is 75 percent and the renal survival--that is, survival with functioning kidneys--58 percent.

[Slide.]

I will point out that the survival curves that I show you, then, that go out to ten years, that is all of the patients going out to ten years. This is not a Kaplan Meier curve where three of your hundred patients make it out to ten years and you run it out there. This is a true ten-year follow up.

During the course of the treatment, 36 patients--42 percent--achieved what we defined as a remission. Now, remission isn't cure and remission isn't regression. Remission is remission.

[Slide.]

The way we defined remission is the serum

creatinine went below 1.4 or stayed below 1.4 and the urine protein decreased below 330 milligrams per day. As you can see, it took a couple of years. Here is the time to remission for these patients to get in.

[Slide.]

The reason why it takes this long for a patient to qualify as having remitted is because it takes that long for the proteinuria to get that low when it starts off at 6 grams.

So that is a consideration for design of the study or for the definition of an endpoint.

[Slide.]

What is the significance of inducing a remission? The reason I show you these data is because it is significant enough, I think, that we have to look at whether we can determine those parameters associated with remission because that might help us both in designing a study and, perhaps, in determining endpoints in the end.

[Slide.]

Here you see it; renal survival. Renal survival means that the patient doesn't go into renal failure. Of the patients who achieve a remission, as you can see, over 80 percent of those patients going out over ten years, continue to have not adequate renal function but essentially normal renal function whereas those patients who don't have

a remission, that is where the renal failure data derives.

So a remission is a critical goal. In terms of patient survival, it is the same story, you see. By remission, I mean renal remission. We are not going to be talking about pleurisy and so forth right now. But if they achieve a renal remission, their survival approximates the survival of lupus patients who never had kidney disease whereas if they don't, that is where your patient drop off can be found.

[Slide.]

The status at final follow up depending on whether they achieved a remission or not you see here. Mean follow up of those who achieved a remission is now 155 months and, of course, it is lower for those who didn't go into remission because many of those patients died or went into renal failure earlier.

As you can see, in terms of dialysis, the great majority of these patients never had a remission. Renal death, the same; relatively rare, actually, if they get a remission. Non-renal death; no significant difference. Stable renal function means that they can still have some proteinuria but they have a serum creatinine of less than 1.4. As you can see, 83 percent of the patients who achieved a remission were there while only 20 percent of those were not.

So remission is a critical goal.

[Slide.]

Whether a patient goes into a remission or not depends a great deal on whether their renal function is preserved when they begin therapy. The histology in these patients was the same. You would not be able to tell which of these groups a patient falls into by just looking at the histology.

However, if they came in with a creatinine of less than 1.4, their creatinine generally went down from where they came in but it tended to stabilize at 1.4 and their proteinuria resolved. Patients who had elevated creatinines, the higher the creatinine, the less likely they were to be able to attain remission status. So I think that tells you something, perhaps, about the window that you might use in terms of developing a study.

[Slide.]

Can long-term prognosis be predicted by the initial response to therapy?

[Slide.]

The plasmapheresis trial, for those of you who know it or remember it, was very rigidly administered. The renal status was looked at at four weeks very carefully and a decision was made if the patient was stabilized or improving, they then went on this taper protocol. If they

got worse, they would be recycled; in other words, they would get another month of Cytoxan and they would stay on their high-dose steroids and they would get another month of plasmapheresis if they were in that group.

As you can see, at four weeks, you are beginning to see a predictive clinical course so that it may not be necessary to follow these patients out for very prolonged periods of time if this holds true. This is 86 patients which, by the way, at least at the time, was more patients than were probably in all the other clinical trials in lupus that had been reported up to that point.

You can see that, of those patients who were worse at four weeks, the majority, although it is not quite statistically significant, were ended up on the non-remission group.

[Slide.]

Does the occurrence of lupus exacerbations predict long-term outcome? This has been discussed. In terms of our own information, we defined all of the extra-renal flares and all of the renal flares that we could think of because these patients were all treated uniformly. So, for each type of flare, they got the exact same amount of prednisone for the exact amount of time, whichever center they were in and so forth.

As you can see, in terms of both minor and major

flares, there was no difference in terms of the remission and no remission group. That is for extra-renal.

[Slide.]

But, in terms of renal flares, mild was going from a normal urinalysis to an active urine sediment. Moderate was the creatinine goes up at least 0.3 or proteinuria increases by a gram. Severe, the creatinine had to go up by at least 1.0. As you can see, of those patients who tended to go into remission, they tended to have, certainly, fewer severe renal flares and, therefore, fewer renal flares but probably not dramatic enough to be able to power a study.

[Slide.]

Does the histologic classification predict the outcome in severe lupus nephritis?

[Slide.]

In terms of renal survival, there were three types of lesions that could get a patient into this definition of severe lupus nephritis. The red line was DPGN, or so-called WHO class 4. That is diffuse, inflammatory glomerulonephritis. They did the best. The blues were segmental glomerulonephritis. This means only a segment of the glomeruli were involved and at least 50 percent of glomeruli. The yellows were membranous which superimposed either of these two lesions.

Interestingly enough, the diffuse proliferatives

actually did better in this study than the other two. I don't think that you can power a study around histology and leave the DPGNs out, for example. But I think it is important to point out when you are comparing studies that histology is a determinant.

This is a surprise, by the way. I think that most people would have said that the results should have been the opposite.

[Slide.]

Are there factors which are predictive of renal remission?

[Slide.]

Looking at our multivariate analysis, no progression at week 4, DPGN as opposed to the other--class 4, in other words, as opposed to 3 and 5. Jim will be glad to see that we actually were able to find the chronicity index being significant. In our previous publications on the negative outcomes, it wasn't. But, in terms of predicting a remission, the more chronic changes in the biopsy, the less likely the patient would go into a remission which is another way of saying that if a patient has protracted the disease, they are not likely to go into a total remission.

Being white was certainly an advantage over other races and the less proteinuria, the more likely they were to

go into remission, the lower creatinine, the more likely they were go to into remission.

[Slide.]

Are there factors which are predictive of end-stage renal disease?

[Slide.]

We found, of course, the creatinine and the remission, as you have heard. The only reason I put anti-Ro up, aside from the fact the it is a risk factor for end-stage renal disease is that it was the only serology, of all the serologies that we ran, and we worked with Mo Reichlin on this as well--of all the complements and the anti-DNAs and the rest, cryoglobulin, C1Q-binding activity, all the rest, this is the only one that showed up as a predictor in any of our models.

[Slide.]

What about the surrogate issues that we have been discussing?

[Slide.]

The doubling of serum creatinine, I don't know. I think that in our diabetes study, the captopril trial, I think that was the first time in a real trial that doubling of serum creatinine was used as an endpoint. I didn't think it was going to become an industry standard. I won't even tell you how we came up with that.

But, clearly, there is a clear relationship to halving of the glomerulofiltration rate, and the advantage of using serum creatinines during a renal study is you get a lot of points on your curve. If you going to do iothalamate GFRs, there is a limit to how often you can do those.

[Slide.]

In the diabetes trial, when somebody halved their iothalamate clearance, and this was 400 patients with type 1 diabetic nephropathy and, I believe, about 60 of time doubled their serum creatinine--at doubling of serum creatinine, the iothalamate clearance actually more than halved to the creatinine and, of course, the BUN doubled, so that there is no doubt that doubling of serum creatinine is a good measure of a major loss of renal function. It is cheap and it is easy to do.

[Slide.]

The advantage in a disease like diabetes, for example, is that the median time from double of serum creatinine to end-stage renal disease was only nine months so that even if someone was applying a very strict criterion to you of having not creatinine, which is considered a surrogate endpoint in some parts of the FDA, but renal failure, it only takes nine months to get from one to the other.

Now, unfortunately, in lupus, I don't think that

is the case. That is the problem, that lupus is an episodic disease. It is not one of these progressive diseases. I don't think you can model a study--you can't power a study around it.

[Slide.]

In terms of what we are talking about using our programs, what we looked at was what would it really take--I think there are a lot of people in this room who want to know whether it is possible to actually study lupus in terms of drug development and so forth.

I really do believe it is. This 30 percent value, you could use it either for positive or negative. About 30 percent of our patients either died or went into renal failure and 30-some percent went into a remission. So either the positive or the negative outcome could be used here.

You can see that, for an 80 percent power study, you need a drug that is really going to work in order to get you into the area that is practical in terms of the number of patients with lupus nephritis that are available within a reasonable period of time.

[Slide.]

I you go up to 90 percent power, you can see the figures changing and your drug clearly has to work and your failure rate has to be substantial. This is not impossible.

It is difficult but it is not impossible. This is doable.

[Slide.]

What I think is not doable is to try to study a drug for equivalence with the attempted claim that it is safer than cyclophosphamide. Probably almost anything being developed by any manufacturer in the room is safer than cyclophosphamide. But the fact of the matter is that you are going to need these kinds of numbers to show a difference if you are just going for equivalence, equivalence outcome, more safe.

In closing, I must say that as difficult as the study of lupus nephritis is, we have done it. It is possible even using difficult endpoints like end-stage renal disease and doubling of serum creatinine and death, or the positive one of remission, it is still doable.

I think there is going to be a need for some give in terms of what is required of a study because it is a lot more difficult to do the study, I can tell you, than to study diabetes which is what I have been studying for the last ten years.

The alternative to not doing these studies is just what we have right now, these dangerous drugs being used off-label in this population of young women. 30 years ago, we had the same discussions at rounds about the use of alkylating agents that we had last week. I think that is

something we have to improve upon.

So I think that, in the end, we have to be able to do these studies. Thank you.

DR. ABRAMSON: Thank you very much, Dr. Lewis.

We will, again, entertain questions when we discuss the questions put to us.

Open Public Hearing

There are no individuals formally registered to speak at the open hearing. Is there anyone present who would like to make a statement at this time?

If not, we will reconvene at a quarter to 3:00.

[Break.]

Panel Discussion

DR. ABRAMSON: We will have discussion of the clinical trials in lupus nephritis and other organ-specific manifestations of SLE. We will try and go through the four questions that were put to us. Many of the issues have been addressed preliminarily by our speakers.

We do also, as we did this morning, have guest expert panelists joining us for the discussion. Actually, I would like them each to introduce themselves and their institution, if they would.

DR. BALOW: I am Jim Balow from NIDDK, NIH.

DR. DONADIO: I am Jim Donadio. I currently do what I wish since I am retired from the practice of medicine

and nephrology at the Mayo Clinic. I am doing some visiting faculty work at the University of Arizona in Tucson which is a good place to be if you are from Minnesota at this time of year.

DR. WEST: I am Sterling West. I am a rheumatologist at the University of Colorado.

DR. ABRAMSON: Thank you very much.

I think, as we did this morning, if the panelists would like to take a couple of minutes just to make a statement about their views, particularly on question No. 1. Why don't we start with question No. 1 and start with your opinions.

DR. BALOW: I am certainly going to echo many of things that I think Dimitrios Boumpas has already put on the table in terms of our views of what we would think are feasible measures of outcome in studies of lupus nephritis.

Clearly, one can take a range of points of view as to what you would like to see in any study in therapy in lupus nephritis. Ed Lewis has shown that it becomes unequivocal if you can show a difference in the risk of end-stage renal disease. Our early studies, certainly, I think, were least controversial by using the end-stage renal failures in outcome.

But, again, as Dimitrios showed, the delay in that outcome and the infrequency of that outcome means that you

have to have huge numbers of patients in those trials. So it seemed a number of groups have actually lit on the idea that doubling of serum creatinine, which means at least a 50 percent reduction in GFR, is a reproducible and a very good predictor of subsequent end-stage renal disease.

So we like it as a measure of outcome. We think it is relatively non-controversial. The doubling, we think, is high enough and, if it is sustained, usually cannot be explained by other intercurrent problems like uncontrolled hypertension, use of ACE inhibitors, use of non-steroidal agents. So those influences don't usually confound the interpretation of that outcome very much.

I think what is even more intriguing to us is whether we can actually accumulate longitudinal data which would suggest that, in a meaningful way, we could light on something that is even more sensitive and, perhaps, an earlier predictor such as a 50 percent rise in serum creatinine.

I think those are data that we need to collect probably by looking at experience in different centers around the country and seeing how well that predicts doubling and how variable it is and how much lead time we could get in predicting the risk of end-stage renal disease.

DR. DONADIO: We are like lawyers in that none of our answers are short, are they? I have a small preamble; I

can say this because I have experienced it and that is, over the past nearly half century, there has been a great improvement in both patient and renal survival apart from immunosuppressive agents in clinical trials.

It relates to many things that I think come with a maturing practice, the ability to handle things like high blood pressure which hasn't even been mentioned in any of this but I think it should be brought up that that has to be exquisitely monitored and controlled; the development of subspecialties in medicine which devote themselves to lupus; earlier diagnosis and milder disease because of the serologies that allow people to diagnose this condition in its earlier forms than when I started out and saw only very late disease.

Concerning the issues of doubling of serum creatinine, it is a conventional endpoint but I think there is some treachery involved in that one is using big medicines and if a doubling of serum creatinine occurs, indicating activity of the renal disease, big medicines are going to be used--and I use that generically to include high-dose prednisone--that will allow that marker to fall to less than a doubling and you may be missing and important endpoint in that improvement.

To say it another way, there isn't a linear progression to renal failure once the serum creatinine

doubles as there is, and has been pointed out, in other primary renal disease or in diabetic nephropathy, for example.

Normalization of an elevated serum creatinine assumes an abnormal baseline serum creatinine and will limit your recruitments so that you must take people who have normal serum creatinine levels who may also have impaired renal function that is hidden in that normal value. Everyone knows the old Doolan curve. Doolan was at the Bethesda Naval Medical Center many years ago and showed this slope of large change in serum creatinine covering a reduced renal function by as much as 50 percent with the creatinine still being normal.

A decrease in the frequency of renal flares has already been pointed out. After big medicines are used for severe disease, it takes a while for a renal flare to occur so the trial must go on for a long time before one sees enough events to capture, if you are comparing treatments and using renal-flare outcome as your measures.

The decrease in proteinuria, I believe, as a secondary endpoint must be linked to changes in renal function for it to make any sense, as has been pointed out by the decreased, so-called fixed proteinuria, as not being necessarily a bad marker. Remission of proteinuria, as Lewis has pointed out, is an important marker and indicator

of the good prognosis.

I have covered No. 1, I think, in my view.

DR. WEST: I think that most of my comments would basically echo what has already been pretty well discussed here. Doubling of serum creatinine and decrease in the frequency of renal flares, if that can be better defined, I think would be the most valid endpoints.

I think, though, that my comments are more for question No. 4 when we get to it.

DR. ABRAMSON: I would open it up to the panel if they have discussions of the speakers or the expert panelists.

DR. FELSON: I'm glad we waited. I was very exercised at the break and I have gotten less exercised now. I must say that whereas this morning we heard about global lupus, a bunch of activity indices where we are desperately in need of clinical-trials data to look at the sensitivity to change to compare those instruments, I was pretty depressed listening to this set of comments this afternoon because, actually, there are a lot of clinical-trial data here from the NIH especially.

The opportunity to look at the relative sensitivity and discriminant validity of any of these definitions or renal deterioration or renal improvement is readily available from those data.

It is striking to me that those data haven't been plumbed to ask and answer these questions. This is an answerable question with the data currently available. Why isn't a smaller increment in creatinine over time likely to be more sensitive to change than a doubling of serum creatinine?

The proposed measure of renal deterioration that Dr. Boumpas suggested about red cells being present or absent, some proteinuria being incorporated, that is readily testable, I think, in the NIH datasets of clinical trials. Why can't we see the results of that stuff? Why can't we know from data? There is multiple-trial data. There is not just one-trial data.

Why can't we know what these data show? It seems like it is an answerable question.

Interestingly, Dr. Lewis, then--I think Dr. Lewis--puts up a slide suggesting that number of patients needed in trials to test out very important treatments for a very important disease are to test, to evaluate or detect an effect of 50 percent which, therapeutically, is a very big effect or 33 percent which many of us would also characterize as a very big effect, are very, very large numbers, larger than any of the NIH trials have had, perhaps as large as the plasmapheresis trials had.

This suggests to me, even more, that this field

would benefit greatly by some kind of analysis which focusses on maximizing the discriminant validity of these outcome measures. My goodness; one could probably reasonably do a trial with many fewer subjects figuring out what works if one simply knew how to measure the outcome in a more efficient way than doubling of serum creatinine which is a threshold which seems quite high.

You figured out yourself, I think, correctly that going from end-stage renal failure to doubling of serum creatinine required many fewer numbers. Going from doubling of serum creatinine to a lower threshold probably requires commensurately fewer numbers.

Using mean serum creatinine in treated groups over time probably requires even fewer numbers because it then allows you to pull in the improvement to some patients who were treated. I know it is a rhetorical question, but it would seem awfully useful for that work to be done. I guess I am not sure why it hasn't been done or whether it has been done and perhaps we can get some insights from it.

DR. BALOW: Thank you for the comments. Basically I think you were the one this morning that said we have to crawl first and then we walk and then we run. But I think we actually went from a study where we used end-stage renal failure as the ultimate outcome to studies that have looked at doubling serum creatinine. I think it worked quite

successfully.

We are actually hoping to show additional follow up in that study in the near future and we want to actually go back and look at the utility of using lower increments in creatinine, as you are suggesting.

DR. LEWIS: I know I risk sounding pedantic, and I am surely not going to defend--I am sitting in between two people from the NIH and I never really believed in their studies so I am not going to say much about this.

But at the risk of sounding pedantic, I have to tell you that you can't do a study by putting a bunch of patients and measuring everything that you want to measure without declaring, at the beginning of your study, what you actually are going to use as your endpoint, analyze things forever until you come up with something positive.

That study has no power. That study is invalid. Now, it may be a valuable thing to do for your next study, but it doesn't help. So I think it would be great, actually, if Jim and Dimitrios have this kind of data because it would help us with the next one down the pike.

But when it comes down to it, there is a science of clinical trials and there are some pretty well-stated biostatistical requirements of a valid clinical trial. And I am all for that.

DR. BOUMPAS: Your point is well taken. We are in

the process of reducing these data. We had to extend the follow up of the patients because the numbers we had, we could not come up with the numbers, analyzing the numbers we had. From the short-term follow up, we could not come up with something meaningful. But these data will be coming in the next two or three months, so your point is well taken.

DR. PETRI: I have two questions for our discussants. One is wouldn't a true measure of GFR be a more sensitive surrogate variable than looking at creatinine, especially technetium DTPA or iothalamate. My second question is why didn't anyone discuss dysmorphic red cells.

DR. DONADIO: Because, as was pointed out, you don't even have to examine the urine sediment to be a reputable clinical lab anymore. The first-morning urinalysis and sediment exams, including dysmorphic red cells, which mean the same thing to me as a red-cell cast, is the most valuable microscopic examination of the kidney. But you can't get it.

All of these trials have to be multicenter. So you cannot, I don't think, regulate how urinalyses are being done. That is why somebody said hematuria is good. I don't think it is any good. I think you have to examine the urine sediment on a first-morning voided sample.

As far as measure of GFR, and the others can speak

to this, that has been done in the so-called end-stage renal-disease MDRD study, the most expensive test of all. It didn't help any more than serial serum creatinines. I may be speaking a little out of school. Very expensive. Again, very hard to examine reproducibly by all labs.

DR. ABRAMSON: I have a question. The doubling of serum creatinine appears to have a consensus. But Dr. Boumpas, you didn't think that a significant reduction in the serum creatinine could be used as an endpoint. Why couldn't a similar kind of parameter be a surrogate, at least as I understand the presentation?

DR. BOUMPAS: Are you referring to the normalization of the serum creatinine?

DR. ABRAMSON: Yes.

DR. BOUMPAS: I discussed two limitations that this test has. I think, as a primary outcome, as I told you in our experience, which is different in many ways from the experience that Mr. Lewis presented, only one-third of the patients had an abnormal serum creatinine, at least in the latest study.

So that limits its utility. If only one-third of your patients have a normal serum creatinine, that is a problem and that will increase the number of patients. The other issue is the change, the relative change is different. I brought up the example that if the creatinine is 2.0 and

normalizes to 1.0, or if it is 1.3 and it goes down to 1.0, I am not sure that you are dealing with the same effect.

So these are the two main limitations.

DR. KATONA: Dr. Lewis, I would like to ask a question from you. You presented data indicating that a little bit less than half of the patients, their lupus nephritis went into remission and, basically, the long-term outcome for them was excellent.

Could you please delineate what therapy were these patients receiving.

DR. LEWIS: I can, to a point. The initial therapy, as I indicated to you, was high-dose prednisone which was either 60 or 80 depending upon size. If, at four weeks and then at a variety of node points along the way, they were stable or doing better, that got tapered to 20 every other day over about a 30-or-so-week period.

The initial cyclophosphamide--and all of the patients got cyclophosphamide because this was plasmapheresis study. The plasmapheresis had to get it so the controls did. They all went on 2 milligrams per kilogram per day of cyclophosphamide for four weeks and one 1 milligram per kilogram for another week. So they got about a gram a week for four weeks and half a gram for the fifth.

So the patients who did well actually then went on

this 20-every-other-day protocol and we had very stringent protocols for each of the problems, complications, that we were expecting so that everyone would be treated the same, the reason being that since most people thought that plasmapheresis would work, we didn't want to compare plasmapheresis to the control group doing worse, getting more prednisone, which might wash out the difference.

As it turned out, that wasn't the case. Patients who had severe exacerbations, severe renal flares or severe extra-renal flares, actually went through that original protocol again, and there were a couple of them who went through a couple of times.

The last five years or so, we haven't kept the protocol going although the group continues, the collaborative study group continues in other areas. So I have maintained a knowledge of what is going on and I don't think anything much has changed.

I think that we were all so happy with the way this protocol worked that, basically, it went into all of our clinical practices.

DR. ABRAMSON: Ask your questions, but try and bring your questions around to the point of this question which is the surrogate markers

DR. FERNANDEZ-MADRID: I wanted to comment on Dr. Boumpas' discussion of potential surrogate markers. I

believe that his discussion was backed up by knowledge of the renal pathology in his cases; is that true?

DR. BOUMPAS: What was the question? Are you referring to the proteinuria?

DR. FERNANDEZ-MADRID: Your conclusions on the doubling of the serum creatinine were based on knowledge of the pathology of the kidney in these patients? Did they have kidney biopsy?

DR. BOUMPAS: Yes. The great majority of our patients have kidney biopsies.

DR. FERNANDEZ-MADRID: For me, it is important. I was surprised by the data of Dr. Lewis that showed that, apparently, proliferative glomerulonephritis did much better than segmental and membranous. At least, this was my perception.

This is not the experience of others. To double the creatinine in a patient with diffuse glomerulonephritis may take a few days or a few weeks, and the evolution of a membranous lesion, the creatinine may double in weeks, years, maybe. So it is a completely different matter.

DR. LEWIS: I think that either I didn't make the point or it may have been missed. In order to get into our study, a patient with segmental glomerulonephritis had to have a clear majority of their glomeruli involved with the process. Most of them, of course, had 60 to 70 percent of

their glomeruli involved with the segmental lesion.

In terms of membranous, in order to get into the study, it had to be membranous plus either segmental affecting more than 50 percent of glomeruli or usually, more often, membranous plus diffuse proliferative glomerulonephritis. So we are not talking about membranous glomerulonephritis here. We are talking about the inflammatory lesion.

DR. FELSON: Let me bring up a point that was, in part, raised by Dr. Boumpas and Dr. Petrie and try to get back to it a little bit which is you commented on improvement being problematic to measure essentially because improvement in certain increments of creatinine didn't necessarily represent the same degree of renal functional improvement.

Michelle asked you about different ways of measuring renal function. Let me also add to that question the idea that a doubling of serum creatinine does not represent necessarily the same degree of loss of renal function at every baseline level of creatinine.

One of the questions, I guess, then, is should we be using creatinine, should we require a fixed increment in creatinine as a measure of improvement or should we use other measures of renal function to try to get at improvement or deterioration--improvement or deterioration,

by the way, because the argument you made holds also for deterioration.

You said, "look, a certain amount of creatinine improvement doesn't signify the same improvement in renal function." That is also true of deterioration. You can make the same argument there.

DR. BOUMPAS: Just one brief comment. If I had the choice, I think the more sophisticated way is to estimate the glomerular-filtration rate. But I am not sure that this is practical for multicenter trials, in my experience. For the small studies that we do at NIH, they are part of our evaluation. But I am not sure how practical they are.

DR. LEWIS: Not practical, I can guarantee you. But the MDRD which had 500-and-some-odd patients, they were doing iothalamate clearance on those patients, I think, every four months. It can be done. I think you have to realize that even in the best of hands, the variance in iothamamate clearance is somewhere between 14 and 18 percent. So, even though it sounds like you are describing this gold standard wonderful test, glomerular-filtration rate, there is a substantial variance in iothalamate clearances.

There is not that variance in a serum creatinine and you get a lot more points on your curve which is what

really gives the power to the curve. So you can do it either way but you ought to realize that, in a multicenter trial, that will involve a number of patients, however many, let's say, it will be between 100 and 200 in a lupus-nephritis trial, that it will be incredibly expensive.

DR. DONADIO: It's true, if I may add to your question, that a doubling of creatinine from 3.5 to 7 is not the same as 1.0 to 2.0, obviously. 3.5 to 7 is end-stage. 1.0 to 2.0, it could be a 50 to 60 percent reduction in GFR but there is still useful GFR.

That is why it is important, I think in any trial that considers lupus nephritis, that the stratum of renal function must be considered. Then you equalize out the baseline creatinines, normal and abnormal, that is a way to stratify or normal to 1.5, 1.5 and higher, so that the changes up or down are equal across both groups with differing baseline levels of renal function by serum creatinine.

DR. ABRAMSON: One more comment on this issue by Dr. Balow.

DR. BALOW: Just a comment on the GFR issues. It is obviously something one would like to have but, in a disease like lupus nephritis in which there are remissions and exacerbations, there might be enough fluctuations in the absolute GFR numbers to make it very difficult to identify

short-term trends. That has always been a bug-a-boo with using either reciprocal serum creatinine as to how to actually draw the slope of change or to use the GFRs, the true GFRs, which are very expensive and actually rely on them to measure before and after.

DR. LOVELL: I have a question about the role of renal biopsy and some of the data you presented depended upon the performance of renal biopsy at the beginning to show prediction of outcome based on renal histology. I base this on the fact that some of the new treatments will hopefully be less toxic than cyclophosphamide and so the compulsion to do a renal biopsy, to say who would be eligible for these potentially less-toxic treatments may not be as compelling as it was when you were talking about potential long-term toxicity from cyclophosphamide.

So I was wondering what the panel thought about the role of renal biopsy in lupus-nephritis trials.

DR. LEWIS: I think every patient with lupus who has any abnormality in their urinalysis whatsoever needs a renal biopsy. That is absolute. I am shocked to hear that it could be different.

DR. BALOW: The analogy of the damage index that we actually have used in the study we published in 1984 was actually looking at pre- and serial renal biopsy and actually trying to measure the acquisition of scarring,

atrophy and fibrosis in the biopsies.

So it was really a pathologic damage index. In fact, it was a powerful predictor of what actually was to eventuate in that patient population in the risk of end-stage renal disease. So it can be used but it is not, again, very practical to consider serial renal biopsies and using pathologic outcomes as a surrogate for later functional outcomes.

DR. SILVERMAN: My question is trying to understand the concept--I am using to creatinines changing with age. As a child grows older, we have different limits for creatinine. I hear repeatedly from the nephrologist that a creatinine of 1.2 may be distinctly abnormal, especially if you are a 40-kilogram thin woman versus a 200-kilogram man.

Why, then, are we hung up on changes. Why don't we go back to percent changes. Dr. Boumpas commented on 1.5 going to 1.0 being different than 2.0 going to 1.0, because there is a significant difference in percent change.

Since we are going to have baseline creatinines on everybody, and I am told creatinines are fairly reliable day-to-day, why don't we look at each patient with percentage change so one can see improvement going from 1.2 to 1.0. I have been convinced that that is significant because most nephrologists would say if you went from 1.0

stable for years to 1.2, that is not good.

So I just don't understand the reluctance to go the reverse.

DR. LEWIS: You won't find any reluctance here. It varies with the disease state for some reason. But I think that serum creatinine is as an extremely sensitive marker for glomerular filtration rate in the given patient. I think the problem is, and there has been a lot of literature written about this, that people have tended to sell the truthful data that if you have a serum creatinine of 1.2 and it is in a 25-year-old football player, that means that that person's glomerular filtration rate is 140. If you have a serum creatinine of 1.2 in some little old lady in a nursing home, that patient's glomerular filtration is 20 so a creatinine of 1.2 doesn't mean anything.

That is not true at all. In a given patient, all other things remaining equal, like they don't take Creatin to hit home runs and things, minor changes, 10 percent changes, certainly, in serum creatinine reflect that much of a change in the glomerular filtration rate.

DR. BALOW: The reason it doesn't work is because the kidney has two ways to actually improve its glomerular filtration rate. One is to actually heal the pathologic injury and the other is to compensate by either hemodynamic changes or hypertrophic adjustments.

If you are trying to tie that improvement of pathophysiology to your intervention, you don't want to be looking at compensatory adjustments that the kidney can make. That is not what you are trying to evaluate. I think that is the most problematic area of looking at percent change in the population.

DR. WHITE: One thing that I was must struck by was Dr. Lewis' data that you could predict those likely to get remission by normal serum creatinine to start with. Those data, to me, would make me think that you would not want to use normalization of elevated serum creatinine as an outcome because that would mean that you would have to exclude people who had normal serum creatinine to start with if that was going to be your outcome. Hence, you would exclude the people most likely to benefit.

So I would like your thoughts on that. I would also like to know what you think needs to be done with people with different levels of serum creatinine at the time of entry.

DR. LEWIS: I think Dr. Boumpas addressed that a bit. In their studies, of course, a much larger proportion of patients had a normal serum creatinine. I think that what we found was in the patients who had a normal serum creatinine and severe glomerulonephritis, that when they responded to treatment, their serum creatinines actually

came down so that the young woman who came in with a serum creatinine of 1.4 probably would end up with a serum creatinine of 1.1 or something of that sort so that you could work out--you could use a delta or a percent decrease and still use patients who had relatively normal glomerular filtration or at least creatinine clearance.

In terms of the business of people coming in with different levels of renal function, I have to say that you have to use a sample size that washes that out. That has been the problem with the great majority of studies of lupus nephritis. There are too few patients in the study and you can get hurt by just a few patients who have a marked decrease, for example, in their GFR.

If you have enough patients in the study, that should all fall out. That particular problem doesn't trouble me at all. I think that while 100, 150 patients in a study sounds daunting, it is certainly doable and it takes a lot of work. I think if that is what it takes to find out whether a drug works in this problem, then that is what it will have to take. I don't think there are going to be many shortcuts here.

DR. ABRAMSON: I think, in view of time, we are going to shortly go on to question No. 2. But I would like to see if we can get a view of consensus among the panelists with respect to the surrogate markers in question No. 1.

Does anyone of the invited panel disagree with doubling of the serum creatinine as a surrogate marker? That seems to be a consensus.

DR. DONADIO: One qualifier. Only if someone can show that the doubling of serum creatinine in previously run trials clearly relates to a much higher proportion of those who go on to end-stage renal disease, so, as an important marker of relentless progressive disease.

DR. LEWIS: It's true in ours, Jim. I showed it. And Dimitrios showed it as well. I just wanted to add something. Dimitrios covered this but the sheet doesn't, and that is--and we found this very useful in both our type-1 and type-2 diabetes studies. There is nothing wrong with composite endpoints. Personally, I feel that doubling of serum creatinine is a valid endpoint in a lupus study. But I think it could easily be doubling of serum creatinine plus renal failure plus death as a composite endpoint which is valid and actually would help the sample size a lot.

DR. HARRIS: But can we press just a little further than ask, okay it is doubling. But what about 50 percent. How is that? Or 30 percent?

DR. LEWIS: I think 50 percent probably is fine. I think we have to show you the data. But I bet the data will show the same thing.

DR. BALOW: I think we need to get that data and

really see how they correlate. I think that is very, very important.

DR. ABRAMSON: There was a view expressed that normalization of the elevated serum creatinine was not a surrogate endpoint. Do we have agreement of the panel on this or not? I am just trying to address the FDA's questions. It is point No. 2, normalization of elevated serum creatinine was not a primary outcome criteria.

DR. BALOW: Again, I would like to just emphasize to the group that I assume it is my experience as well that the number of times that one can see normal GFRs in patients who got 40 to 60 percent of nephrons that have been totally and permanently destroyed is rather striking. So I, again, come back to the point that it isn't that you can only normalize serum creatinine by regression of pathology but actually by compensation. So that is why I don't like it.

DR. LEWIS: I think that the remission data actually speaks for itself. I think we have to be careful, and I think over many years Jim and I have been unable to agree, at least in part because of a difference in the populations that we see which Dr. Boumpas actually referred to.

I think that when a patient with severe lupus nephritis comes to rush, they come to rush and they may be in what before HMOs were our traditional referral lines but

they are patients in the wild which is actually the kind of patient that a study should be directed against.

I don't know this for a fact but I have always imagined that the patient who comes to Jim for one of his studies has been to one of our referral people, has been to us, has been to Mayo's and has finally ended up on Jim's doorstep. I think that is a somewhat different patient.

I believe that these so-called remission kind of criteria should be seriously considered. They occur as often as the negative things like renal failure and death so why not use that. Why not use a positive endpoint rather than a negative endpoint.

DR. ABRAMSON: Let me go on in view of the time. The proteinuria question, the decrease in proteinuria not as a surrogate-outcome measure. The obvious question that would come up is the membranous nephritis patient with normal renal functions.

We have heard mostly that decreases in proteinuria, a decrease is not a surrogate endpoint from what has been presented.

DR. DONADIO: A pure class 5A which is membranous disease similar to idiopathic membranous disease does carry a different prognosis than the Lewis 5C and D which have a proliferative element and, as we saw, they do badly. So reduction in proteinuria in a pure membranous, which is an

unusual subset, is a good sign and a laudable one in any trial that directs itself at that, to answer your question.

DR. BALOW: I would make the additional comment that I think, too, we could use some collective data on whether or not changes in proteinuria would actually be a good surrogate marker for a subsequent adverse clinical outcome in proliferative lupus nephritis as well.

Some of us suspect that it may be but the data are not there on which to make a decision right now.

DR. ABRAMSON: I think we need to move on. No. 2; could a decrease in the use of toxic agents such as high doses of corticosteroids and cyclophosphamide serve as a valid outcome measure in a trial of lupus nephritis.

DR. PETRI: I am going to answer no for two reasons. One is hopefully there was a reason that the patient was put on the high-dose steroids and Cytoxan and we need to quantify and identify the reason because that is hopefully one of our surrogate variables.

My second reason for saying no is we can't get any two rheumatologists to do it the same way.

DR. SILVERMAN: I agree with the last statement, but I applaud Dr. Lewis' study. I think we have to use that as the gold standard. Now, nephrologists are probably just more open-minded than rheumatologists are. Certainly, that is my experience. But if they could do it, I challenge the

people at this table to say, "Why can't we be less pig-headed and follow a protocol?"

One could than say failure to follow that protocol at certain times, it's a flare, et cetera, et cetera, so you have the option. So I am not as pessimistic as Dr. Petri is.

If I could make one final comment. My experience in a referral center that sees most patients in a large--differs immensely from the NIH. I just want to echo to take patients who are fresh, not seen by--with normal creatinine and to compare them to the NIH group is apples and oranges. I just have difficulty extrapolating too much from that data.

DR. ABRAMSON: Any other comments on question No. 2?

DR. DONADIO: To reduce toxic agents, require that you have severe disease to begin with that require the use of these at randomized. And a long enough follow up would be required so that there could be reduction of these agents over the time.

Remember, in the NIH trials, the immunosuppressive agents were used from four to seven years and many of the toxic complications, including malignancy, opportunistic infection occurred months to years after the start of therapy, not just gonadal dysfunction which we know relates

to cyclophosphamide.

So two requirements; severe disease and long enough follow up to observe a reduction in toxic or adverse events, severe ones.

DR. HARRIS: Just so that there will be at least a 2 to 1 vote that I agree absolutely with Michelle that it is very difficult, I think, for numbers of rheumatologists to come to any consensus as to when to lower prednisone and Cytoxan. Nephrologists, nevertheless, think we can't do it.

DR. LEWIS: I just want to tell you that we spent endless hours in a locked room fighting among one another before we came up with our protocol and lots of people walked out. The reason that the collaborative study group has been successful for twenty years now is that we happen to be a group of people who could say pretty much anything to one another and get away with it. That is the way you are going to have to do it.

DR. ABRAMSON: But what is your opinion about these cytotoxics as endpoints?

DR. LEWIS: I think, as I showed you, it is a sample-size issue. If it is a matter of finding one thing more toxic than the other, it is going to take an enormous number of patients. I don't think that is the way to go. It is an inviting way to go because it would be great to find something less toxic than Cytoxan and just replace

Cytoxan, but I think the n would be too high.

DR. ABRAMSON: Any other comments?

DR. BALOW: I assume that question No. 2 and question No. 3 are inextricably linked. Are you making the assumption that the control group would actually contain these so-called toxic agents and that the experimental therapy that you would be looking at would potentially be found beneficial because of the ability to stop the toxic agents, or am I misunderstanding the two questions.

DR. ABRAMSON: I don't know, in No. 3, that there is any presumption of what the control group would be on in terms of other cytotoxics.

DR. BALOW: Then what are we to understand about question 2; under what circumstances would we be trying to avoid the toxic agents?

DR. ABRAMSON: Can we get some help from the FDA staff members?

DR. SIEGEL: Maybe I can clarify. What we had in mind in question 2 about using an endpoint of decreasing use of toxic agents would be, I think, what Michelle was talking about that you had a protocol for when Cytoxan or high-dose prednisone would be used. The patients would be followed along and these agents would be used if they met the clinical criteria.

Then the endpoint, at the end of the trial, would

be the cumulative use of those agents. Efficacy would be presumed, if the patient is on the new agent, presumably double-blind, those patients required less of those toxic agents because they met the criteria for requiring them less frequently.

In No. 3, the idea is that you have patients who are randomized to receive the new agent or placebo, and the placebo patients need to be given some standard of care. The question is what standard of care should there be in a trial like that.

Does that clarify it?

DR. BALOW: I think so. Again, I would vote no, that this is not a good measure of efficacy. Again, I can cite the plaquenil controversy that is so often brought into the clinical setting where some people feel that that plaquenil can never be stopped once it is started and others feel that, "Hey; try them."

DR. ABRAMSON: Let's go to No. 3; what would be an acceptable control regimen in a trial of renal lupus.

DR. DONADIO: I will hit the ball. I think prednisone is fundamental. I will define severe disease without a renal biopsy as an elevated creatinine, a hypertensive patient with nephrotic-range proteinuria who is also anemic. These are four prognostic indicators at time 0, patient entering a study that indicated poor

prognosis.

As most of you know, there are still many thoughtful nephrologists out there both on this and the other side of the Atlantic who think that further randomized trials are still needed with the old standby immunosuppressives that we have been talking about for thirty years, azathioprine and cyclophosphamide.

DR. ABRAMSON: I guess that is the corollary of this question, isn't it, can you do a new drug trial in lupus nephritis without using Cytoxan as a comparator.

DR. DONADIO: I still believe that you can.

DR. PETRI: I would like to divide this question into two because I think we are at the point where we think about induction trials and maintenance trials in lupus nephritis. I think the induction trials might be a lot shorter than maintenance trials and so your choice of a control regimen might differ whether you are looking at induction or maintenance.

DR. DONADIO: The Dutch are doing that right now, as you all know. They are using intravenous methylprednisolone at the outset and I think a short--a few weeks, up to 12, of cyclophosphamide. Then the maintenance therapy is azathioprine against nothing. I don't know how far along they are, but that is being done.

DR. FERNANDEZ-MADRID: I think this question has

ethical and legal implications and I think the panel should answer whether state-of-the-art therapy could be not used in some of these patients as a control.

DR. BALOW: I don't know what the correct answer is but this is my perspective on it. I think if a patient like Jim has described which I suspect, on kidney biopsy, would probably have crescence and necrosis in a fair number of glomeruli, I think it would be difficult for a lot of people to accept that patient not receiving aggressive cyclophosphamide therapy.

On the other hand, perhaps the more commonly manifesting diffuse proliferative patient will have a more moderate disease expression and, if they didn't have crescence and necrosis, I would have no problem with the trial of steroids or some other agent other than Cytoxan as a control group.

DR. LEWIS: I think from a practical point of view, IRBs are not going to accept prednisone arms in this country. I think that we have to look at whether a study can actually ever get done. I don't agree with this. I strongly disagree with it. I really think that the evidence that alkylating agents add much is marginal, at best.

But the fact is that in this country, prejudices being what they are on the basis of what has been written and so forth, that the IRBs will not allow a prednisone-only

arm.

DR. HARRIS: I believe that renal lupus is not all the same and at least we should do membranous and proliferative differently. If it is membranous, I could see a placebo arm, possibly. If it is proliferative, then I think there is a problem with the placebo arm.

DR. DONADIO: Placebo to prednisone even?

DR. HARRIS: Prednisone I am calling placebo.

DR. DONADIO: The data in the lupus-nephritis trials show that, after six months, 85 percent of patients with diffuse proliferative glomerulonephritis, variable degrees of proteinuria renal function, blood pressure, improve, whether it is prednisone alone or prednisone and an immunosuppressive agent. So this is the induction period.

Now, treatment could be incrementalized after that if, after six months, there wasn't a good response, and I am really talking off the top of my head here, then there could be randomization to one of the standbys, azathioprine or cyclophosphamide against a new therapy.

I am just speculating but I think prednisone alone would fly with certain IRBs with convincing background information. Yes, the patient that I described that Jim Balow just reiterated with all four bad markers, probably none of my colleagues, who I don't even look over any more, we would just use prednisone alone.

But my point is that there is some much diversity in how this is being managed that I think it would be difficult to come up with a control group that would satisfy all.

DR. ABRAMSON: Can I ask a question of the panel, just a show of hands on two questions, two individual questions. Number one, in a drug designed for lupus nephritis, do all patients need a biopsy for entry into the study? Number two, can you do a control that does not include Cytoxan with a new comparator drug?

So just for a show of hands. Who agrees that a biopsy is a requisite procedure prior to entry into a lupus-nephritis trial?

[Show of hands.]

DR. ABRAMSON: Thank you. Just nephrologists.

[Show of hands.]

DR. ABRAMSON: Rheumatologists can vote, too.

DR. PETRI: But, Steve, you didn't say within what period of time before the trial starts. That is what is so crucial here.

DR. ABRAMSON: Do you want to make a proposal? You are starting a new drug trial. You have drug X and you want to--

DR. PETRI: This is a hot topic. It is being debated right now about lupus-nephritis trials. Is a biopsy

within one year of entering a clinical lupus-nephritis trial sufficient. So let me ask the discussants.

DR. LEWIS: In the plasmapheresis trial, I think we gave them six months.

DR. DONADIO: I think that is being pretty liberal. I think it should be within one or two months because if a severe lesion is found a year ahead of time, certainly, there is going to be intervention which will modify that lesion and it can't be used as part of the randomization criteria, I don't believe.

Look at the intervention. We are not just letting that diffuse proliferative glomerulonephritis stand. It is being modified in any competent physician's hands who has that information with that patient.

DR. ABRAMSON: The second question; does the control group have to have an immunosuppressive arm, particularly, let's say, Cytoxan. I guess all who agree that Cytoxan must be included in the control group, I would like to see a show of hands.

We are talking about control groups. We are talking about, let's say, proliferative nephritis.

DR. BALOW: I think you have to subdivide the severity of the presentation. If it is a patient with very severe disease, nobody is going to be comfortable without. But if it is an average, mild, moderate clinical disease

expression or mild, moderate biopsy, then I think there is no problem.

DR. ABRAMSON: Does Dr. Lewis agree with that, because I thought you were saying something a little different.

DR. LEWIS: I think, in my world, people get histologic diagnoses. The histologic diagnosis, if it is severe lupus nephritis--that is, DPGN or whatever--it is DPGN, mild, moderate, whatever else.

So the way I would word it is, I guess, who thinks that the control arm should be high-dose prednisone alone. I think that is the other way of wording it. Biopsy-proven relates to serious disease.

DR. FELSON: Let me offer or suggest a different approach. I think the reason we are all struggling here is we are not sure what the comparator group should get because we know there is evidence suggesting it would be unethical to not allow comparator groups to get certain things. I think that is a very important concept.

So a design that one might adopt, one design you talked about earlier, is an equivalence trial. That is an extreme way of dealing with that problem. Another way to do it is to look at the marginal efficacy of a new treatment on top of accepted current treatment.

So what would happen would be patients would get

what they are going to get, what the clinician deems is appropriate treatment; prednisone in some cases may be appropriate treatment. Prednisone and Cytoxan may be appropriate treatment. Prednisone and Inuran may be appropriate treatment.

On top of that, they get randomized to either the new treatment or the placebo new treatment. That is the way it would need to work if you have already got something that needs to be base therapy, basically. Otherwise, it is unethical.

DR. DONADIO: The problem with that is what are you measuring for an outcome. You are stretching smooth muscle to its maximum. You can't stretch it any more with that little additive drug. You have got to have something in mind, I think, for a better outcome if you are using what is currently thought to be maximal therapy.

DR. ABRAMSON: Let me just frame a question for discussion. To the extent that prednisone plus Cytoxan is now a commonly accepted treatment for diffuse proliferative nephritis, if a new drug were to be brought for testing, could that drug be put head-to-head against prednisone plus Cytoxan making it prednisone plus drug X, or would your IRB have a problem accepting that?

DR. LEWIS: I think that is the one that IRBs would accept. It would be if the control arm was just

prednisone, I think there would be problems. Or if there was a third arm that was just prednisone, I think that there would be problems.

There may be ways around this, but I think, from a practical point of view, my opinion is that an alkylating agent has to be in the control arm.

DR. BOUMPAS: I would like to add to what Jim said earlier. I would not have any problems with including steroids as a control group if you exclude the patient with severe nephritis defined as Jim defined them. If that population of proliferative nephritis, I don't have any problems and I don't think a lot of people would have any problem with the steroids.

If they have severe nephritis, in that population, you have to include Cytoxan, or a cytotoxic regimen.

DR. ABRAMSON: So your control group does not necessarily have to include Cytoxan.

DR. BOUMPAS: Correct.

DR. BALOW: The only additional comment I would have is that, from an ethical standpoint and from just a practical IRB perspective, it depends on how carefully you set the escape valves. This is really critically important. If there are plenty of built-in safety features for the patient to escape with predefined criteria, then the committees are often reassured.

DR. ABRAMSON: Any comments from the committee?

DR. SILVERMAN: I think, coming from a very conservative IRB, I still think I would have no problems, as long as there was the escape valve of having prednisone only and, of course, the exclusion criteria of severe. So I have argued a long time that you have to have--that it is a much longer disease in children, I still think I would have no problem, especially when the NIH group would now say that it is agreeable to them.

With that behind me, I don't see any problem.

DR. LEWIS: There is a problem because you can't be a little bit pregnant. There can't be an escape valve here because this is a controlled, blinded trial; right? So a patient who isn't doing so well, you know you can't, then, give--well, yeah. If the experimental group is prednisone plus some immunosuppressant versus X, versus prednisone alone but you have an option of adding Cytoxan, somebody who is not doing well on the prednisone-plus-immunosuppressive group would then get a second immunosuppressive.

Where is the science there? The other thing I just want to say is that IRB or no IRB, it seems to me that it always comes back to me that I have to be responsible for the sample size. I don't care what anybody here says about what their IRBs will do, I know what will happen when I come looking for patients.

DR. BALOW: I think there is no problem. You have to set the escape valves to be exactly parallel in both your treatment arms. As long as you set the rules to be the same for governing all patients, you can actually use the number of escape events as an outcome.

DR. LEWIS: Because then, if your experiment drug is no better than placebo, then you are doing prednisone plus Cytoxan versus prednisone plus Cytoxan, basically, maybe. I don't know. I think that has to be run through the biostatistician.

DR. ABRAMSON: I think we are going to go on to question 4.

DR. SCHWIETERMAN: Dr. Balow, I think, made a good point. But if you could clarify a little bit more about the details of the escape valve since, if you were to design the study whereby a patient would meet that particular failure criterion, that would become the de facto endpoint by which they would then get an alternative treatment.

So it is rather important, I think, that agency have a rather clear idea about what you would consider failure. Also, you have to keep in mind about the possible delayed effects from other immunosuppressive therapies.

I don't want to necessarily get an absolute from you, but can you give me a general idea about the kinds of early escapes you were thinking about?

DR. BALOW: The most obvious one would be a significant and sustained trend in renal dysfunction. In other words, if the creatinine starts rising and it rises steadily without another explanation of some other intercurrent event, some predefined level of rise and nonreversibility could potentially represent a determination of that patient's participation.

DR. DONADIO: I would predict, based on what we know already in treating diffuse proliferative glomerulonephritis severe that that would be occurring in about 15 percent of the patients. Not a lot. That is a biased opinion, but I still think it would be safe for the vast majority of patients.

DR. ABRAMSON: Any other questions before we leave the renal component? When all is said and done, does it seem that a doubling of the serum creatinine that is the one benchmark that we have all agreed upon today, that is the standard against which the drugs would be judged?

DR. BALOW: I think if we are really limiting our comments to large, randomized, controlled trials, that is the one that we all feel the most comfortable with now. We still like to imagine there might be better candidate markers, but I think for less than randomized, controlled trials, early pilot studies, I don't think that we would advise resting all of our money on that late outcome.

DR. ABRAMSON: Very good. We will do the rest of lupus in a half hour. Discuss the design of clinical trials for other manifestations of lupus including CNS lupus, antiphospholipid-antibody syndrome, cutaneous lupus, fatigue, arthralgias, malaise.

Dr. West, do you want to start off?

DR. WEST: I have been studying central-nervous-system lupus in particular for probably the last fifteen years. At the present time, we have no accepted classification, although that may change. We have very little insight into the pathogenesis of the disease. We have no single diagnostic test that can help us make the diagnosis in all cases. And we have no treatment guidelines.

So, other than that, we have quite a bit of knowledge about CNS lupus.

So I think that if we think designing clinical trial for lupus nephritis is difficult, we haven't seen anything when it comes to CNS lupus. The reason I say that is that whereas there are probably two or three important histologic lesions in lupus nephritis that we have to deal with, there are at least nine or ten different presentations that are generally agreed upon in central-nervous-system lupus that would have to be addressed and that is if everyone could agree on the classification.

Trying to come up with what indicates an improvement or a disimprovement in those presentations I think would be particularly difficult. Coming up with surrogate markers would be, I think, extremely difficult. And then to come up with what the control regimen would be for treatment would hard to get consensus agreement upon.

So I think we are with Dr. Liang's group working on the classification which will soon be published. That is a start, but we have a long way to go before we are going to be able to come up with exactly the types of markers that you are going to need in order to determine whether certain medications are beneficial or not.

DR. ABRAMSON: Do you want to take a stab at how long a study should be and, given the state of the art, what the endpoints, the objective endpoints, or measurable endpoints, are that you would list in a study?

DR. WEST: I think I can maybe tell some of the difficulties with that because if we look at, let's say, a one-year study, the chances of recurrence--most people will either respond or stabilize to prednisone alone. Then what we have to look at is, as you are tapering the prednisone, the chance of recurrence.

And then the chance of recurrence is only about 30 percent. Fortunately, they recur in very similar patterns to what they presented with, but they don't always.

So then the problem is if you present with encephalopathy and then your next manifestation is stroke, how do you measure that as far as response to therapy because the therapy may have worked for your encephalopathy, but it may not have done anything to prevent stroke. So I think that would be particularly difficult.

There are neurologists in different presentations--they actually have scales that they use that we could, perhaps, start by borrowing some, for demyelating diseases and dementia and stroke and the like. That could be a start, but we would almost need to come up with pure patient populations of psychosis or stroke manifestations or the like and then see what the intervention did for that particular presentation.

If we are talking about a couple of hundred people for lupus nephritis, in ten years, at one institution, I have been able to get probably 150 people with a variety of different CNS presentations. It would probably take every academic center in the world agreeing to do a multicenter trial to come up with the power to look at any individual presentation.

DR. ABRAMSON: I guess the corollary is that the presentations may reflect different events, so the that for one may have no effect on the treatment for the other.

DR. FELSON: Quick suggestions about this

difficult. One is that an n of 1 or multiple crossover trials, given the transient nature of some of these effects, may not be an unreasonable approach which would afford you much greater power and much more efficiency in using individual patients.

The other is you said there are no surrogate endpoints, but, in fact, MRI might be a surrogate endpoint for some of these. In myositis, we are using MRI, T2-weighted images, to look at inflammation over time. One wonders whether that correlates well enough with some functional outcomes and with activity of disease to use it.

DR. WEST: Along those lines, certainly n of 1 is something that bears another look. As far as the MRIs, those are particularly difficult, at least in the standard techniques that we presently have. The new techniques that are being looked at as far as T2 decrement and the like may actually hold some value there. But as far as looking at white-matter lesions and accumulating those, there are a lot of different things which can cause that that are not central-nervous-system lupus such as hypertension, migraines, cardiovascular disease, a whole variety of them.

The more of those risk factors you put together, the more likely you are to have these "unidentified" bright objects. So to decide whether one of them is due to CNS lupus or one of them is due to one of these other

confounding factors I think would be particularly difficult.

DR. ABRAMSON: I am wondering if Jeff could give us some idea of what you would like to glean from this portion of the discussion because it is, obviously, very complicated.

DR. SCHWIETERMAN: It is quite complicated and I appreciate those comments, Dr. West. I guess one of the thinkings we had in the group was ways to increase the power of studies using composite indices even if they weren't primary indices but, rather, secondary. We are faced with this problem not uncommonly in biologics with relatively rare diseases where the clinical outcome data are relatively scant and we are forced to go to some extremes. Admittedly, you have to stretch the rules a little about definitions and so forth.

But what I was hoping to get feedback on, frankly, whether in any guise you could envision a CNS outcome event, say, perhaps, coupled with a nephritis event, say, perhaps, coupled with a pruritic event and so forth, by which you would use this as a secondary corroborative measure of what you think you are seeing with the primary endpoint.

I am of the opinion, frankly, that the secondary endpoints in these particular studies are going to be as important as the primary, given the few patients and so forth. But it is not clear to me that we are even there yet

to include it as a secondary endpoint.

Perhaps you could comment on that.

DR. WEST: I am not sure we can include it as a secondary endpoint. Again, we don't understand the pathogenesis well enough but as a secondary endpoint, in and of itself, outside of the activity index that we have already had discussions on. So it certainly would be possible in the context that you are looking at all the things included in the activity index to see if something gets worse or something gets better. That certainly is possible.

In other words, a medication that you are looking at for lupus nephritis helped the nephritis but the person had five seizures and two strokes during that time, you would be concerned as far as its value in lupus overall.

DR. SCHWIETERMAN: How about something like hospitalization for a CNS adverse event with wide latitude given to the definitions. This is something we use with RSV in pediatric disease. You don't necessarily ask that the child has a decreased PO2. You just ask that they have some reason to put them in the hospital for that reason.

I guess what I am looking for is variance around the definition of CNS lupus.

DR. WEST: I think you touch on an important point. A lot of people have this mild cognitive dysfunction

that people are still wrestling with and is that, or small deteriorations of that, considered important. I think that could be argued for an awful long time.

On the other hand, to have a significant event, whether that is encephalopathy or psychosis, transverse myelitis, those events are likely to be severe enough that it would require hospitalization. So, certainly, I think all the things that are listed, again, on the activity index is either any of the ones that are out there would require hospitalization because of the severity of the presentation.

DR. ABRAMSON: Let's move on to the antiphospholipid-antibody syndrome; discuss the design of clinical trials for other specific manifestations. We have an expert, the lone remaining expert. Dr. Harris?

DR. HARRIS: As far as the antiphospholipid-antibody syndrome goes, I think the important point to make is that this may well, certainly, in many patients be a different disorder from lupus. It may be related, certainly part of the family. But it often bears no relationship to lupus activity.

The second point to make is that, of course, with respect to the syndrome, we are pretty far behind lupus. We are just about coming up with preliminary classification criteria about which there is consensus, and that is going to be published shortly.

The next point is that there are three real elements to any classification. One is these patients--the big things are thrombosis, pregnancy loss and serological tests. As far as the serological tests go, there are many of them. Nobody believes, I think, or can show that the serological tests are absolutely related to clinical outcome. So I don't think one would say to look at the serological tests as a marker, necessarily, of disease.

With respect to the two clinical items, thrombosis and pregnancy loss, dealing first with pregnancy loss, there have been at least two more or less randomized clinical trials that have compared prednisone versus heparin and baby aspirin. The outcome there is usually birth of a viable infant.

Now, there are, obviously, a number of other things that one should bear in mind. For instance, there is prematurity. There is preeclampsia and so on. So there are secondary measures, but the primary measure there would be birth of a viable infant. So I think it can be defined to a degree.

With respect to thrombosis, thrombosis is venous or arterial. One would want to separate venous from arterial. Some patients seen to get only venous, many of them, and some only get arterial. Obviously, what one would want to measure is frequency of recurrence and, indeed,

there are a number of retrospective studies, at least, that have looked at recurrence, frequency and the influence of agents to prevent recurrence.

The trouble with any trial is that patients, the episodes of thrombosis are episodic and they can occur as far apart as four or five years. So any trial, in fact, would need to go on for a pretty long period of time.

The next point to make is that many of us now believe, based on the retrospective studies, nevertheless, that Coumadin is effective. Indeed, if you are going to go forward with a trial of a new agent, it would be very difficult--or one would want to use Coumadin plus or minus aspirin in the control arm.

The role for placebo with respect to thrombosis and presentation of recurrence, I think there are very few people that would want to do that.

I think those are the comments that I would like to make.

DR. ABRAMSON: Thank you, Nigel.

I think, as a general statement for this whole area of question 4, each of these syndromes has endpoints that are obviously very specific or characteristic. Obviously, any trial design needs to be geared to them. In the case of CNS lupus, it is four or five or six different diseases with probable different pathogenesis, some of them

steroid-responsive or antiinflammatory-responsive and some not.

I think that one has to do a little bit of categorization. I guess that is what Dr. West and he says Dr. Liang are working together trying to do with CNS lupus to sort of do more carefully a categorization of those things that we loosely call CNS lupus, for example, and divide them into syndromes and think about some of these syndromes as being prednisone-responsive and not-prednisone-responsive.

The antiphospholipid-antibody syndrome characteristically is not a prednisone-responsive disease so that one needs to sort of break these out and then determine what kinds of measurements you are going to do.

But I don't think, unless there are other areas--I think that kind of broad statement would apply in each of these. They need to be fleshed out more.

Are there other comments that people would make or other questions that the agency has? Have we covered most of the questions?

DR. SCHWIETERMAN: Yes, you have. This has been quite helpful. I think that we have enough, plenty, to go back with and write a document that then we would bring back to this committee at some stage for further more detailed commentary. But this has been very helpful. Thank you.

DR. ABRAMSON: I guess, on the agenda, I am supposed to make a closing statement. It will be very brief. I would just say that it was remarkable to sit here for all these hours and listen to the amount of expertise and years of study of this disease that was gathered in this room.

I think that there is a lot of information that is there that now needs to be culled together. I think the indices, the various disease-activity indices, although there may be some differences, the similarities are so great that I think we are very close to having some tools to measure this disease.

I think it is still unclear, in looking for consensus, how valuable measuring a heterogenous disease, the way we do ACR 20s and 30s, how valuable that ultimately is and how this is, in many of our views, an organ-centered disease and you have to think about its treatment in that regard.

The only practical issue that that brings up where some of the panel has some concern about if you have a drug targeted to an organ, what is the impact on the cost and design to make people go through filling out all of these disease activity indices and what not.

But I think there is a good consensus that we should collect as much data as possible. I am not sure we

really came to a consensus about whether this is a disease that you can measure the way you measure rheumatoid arthritis in that respect.

DR. SCHWIETERMAN: I agree with you, Dr. Abramson. We will certainly take all of that into consideration. I fully appreciate the need here for a different approach, in many ways, to the ACR and the rheumatology guidance document given those considerations.

DR. ABRAMSON: Right. Finally, I guess the afternoon was also very instructive. I guess we need something beyond creatinine eventually to measure response to lupus nephritis.

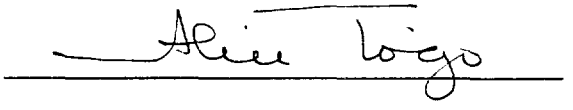
I want to thank everybody for their participation. Thank you very much. The meeting is adjourned.

[Whereupon, at 4:10 p.m., the meeting was adjourned.]

- - -

CERTIFICATE

I, **ALICE TOIGO,** the Official Court Reporter for Miller Reporting Company, Inc., hereby certify that I recorded the foregoing proceedings; that the proceedings have been reduced to typewriting by me, or under my direction and that the foregoing transcript is a correct and accurate record of the proceedings to the best of my knowledge, ability and belief.



ALICE TOIGO