problem of what additional burden do you add if you say you want to see the whole dose range for the response. There is the adverse effect stuff. There is certainly the problem with no placebo group, how are you going to tell what the adverse effect is that's related to the drug.

I just wanted you to get a feeling for the fact that if you wanted to say I really want to know the dose-response curve, that you're not telling people they have to do another 3,000 patients. That's all I wanted to do.

DR. KONSTAM: Can I just ask another question about it? Let's say you do this and you still haven't found an Emax.

DR. LIPICKY: Yes. Well, then you drop back to that simulation business and you top load the top arm because then you have most of your data in the higher doses, which is simulation 4, and you say, I'm going to analyze prospectively. You say, I'm going to analyze this data with the Emax model. You get your estimate of the Emax, your point estimate, and the standard error and 95 percent confidence limit, and I will believe it.

DR. LINDENFELD: Milton?

DR. PACKER: JoAnn?

DR. LINDENFELD: Just for clarification, I have a dose-response curve, and then I have an individual patient. How will I know that there might be specific

covariates that would affect the dose-response curve --

DR. LIPICKY: Well, Lew is going to tell you about that. He says no.

(Laughter.)

DR. LIPICKY: He wasn't prepared to do that.

This won't tell you anything about that, unfortunately. Well, it could. So, he'll tell you then. Don Rubin will tell you.

DR. PACKER: Let's move on. Does anyone have a specific clarification from Ray's supplemental presentation?

If not, let's go on to Lew's presentation. Ray, my sense is that it will take both Lew's and Don's presentations before we can get into the meat of the matter. Is that correct?

DR. LIPICKY: Probably true.

DR. PACKER: So, let's go on to Lew's presentation and see if there are any issues that come up here.

DR. SHEINER: Well, thank you very much. A lot has been promised about what I'm going to say, and I'm not going to actually give you a technical talk. Part of the reason I'm not going to give you a technical talk is in one of my slides, because I really do think we need to keep these things separate. So, I'm taking the title that Ray

gave me, which is what do we want to know, and I'm going to present a somewhat theoretical point of view and with an example. But in trying to clarify exactly what are the issues, the assumption I'm making, of course, is that there is something that we want to know.

I just wanted to start out with a slide referring to a paper I wrote a number of years ago, which tries to say about the whole drug development process exactly what Tom was saying, that there are things that we have in exploratory mode or what I call here learning, and there are things that we want to confirm, and that the entire drug development process can be seen as two cycles of learn/confirm, one early one that gives you the confidence to go ahead and spend all the money that you're going to spend in the later one. In this process, phase IIb and III, as everyone has said, is the phase in which one is mostly interested for our purposes today in figuring out the optimal use of the drug.

Now, just as Tom said, there is no reason why you can't learn while confirming. I actually had that over here. It requires a different view and possibly different portions of the data, but there is generally no reason why you can't do both and that's probably the most efficient way to do things.

But the point that I wanted to make about this

slide is that exploratory analysis, learning as opposed to confirming, confirming meaning I have something that I believe is true and I want to test it, learning meaning I want to understand a whole set of relationships -- that this learning activity is quite different and is what can be applicable to the dose finding. Let me elaborate on that somewhat.

First, let me get these three questions. As I said before, there isn't one. There are three. The first is, what do you want to know? Maybe Don will expand on this more. That's not data. That is to say, the answer to that question is something that exists in your mind in principle that you use data to try to learn about, but you don't want to know the average in this group of people. What you want to know is something about the underlying mean of the population perhaps or something like that. So, what do you want to know is a theoretical construct that you need to figure out what that is.

Then you need to know how certain you need to be. This is where, again, what Tom was saying earlier comes in. There's a tradeoff between certainty and assumptions, and that's the next question, which is what are you willing to assume. What do you think you know already on which you're going to base your data gathering to learn some more? It turns out that questions 2 and 3

have a very, very strong interaction, which I will elaborate on more.

I want to make a couple of points about these three questions, though. These are questions that have to be answered, by domain-specific experts, by regulators, by physicians, by clinicians. Technical people, people who know how to run clinical trials, people who know how to analyze clinical trials really have very little to contribute to the answers to these questions. Mainly the contribution, for example, of the statistician to these questions is to tell you when you have stated how certain you need to be and what you're willing to assume, that that will take you 40,000 patients and 50 years, and you might want to think about that again. But the technical people who are not the domain experts don't really have anything to say about these fundamental questions that are scientific or clinical questions.

The second is that I said before. The second and third questions interact strongly. The more you're willing to assume, the less certain in a sense you can be unless those assumptions are rock solid. But if those assumptions are open to question and if your conclusions are going to depend on them, if you're using those assumptions to gather less data or more specific data and not gather other data, then clearly, if the assumptions are

wrong, the conclusions may be wrong.

Once you answer these questions, to turn the statement I made a moment ago around, study design and data analysis decisions are purely technical within technical constraints. That is to say, that's when you bring in your experts on clinical trials. That's when you bring in your statisticians. With all due respect to Ray and what he's saying, a number of the things that Ray has been saying have been these technical issues. What I want to assure you of is that there are plenty of technologies around that will allow us, when you give us a statement about how certain you need to be and what you're willing to assume, to be able to minimize the cost of learning those things with that degree of certainty. And there are ways to do that that I think we need to understand only because we need to have some confidence in them. But they're fundamentally technical questions that come to us from other experts.

So, I don't think, for example, that the domain expert should be someone who says, I want to use a t test or I want to use a model-based analysis or I want the design to look like this. That's not actually a domain expert question. These are the domain expert questions, and how you get the answers come to us through people who are experts in answering those kinds of things.

So, what are the answers to those questions for dose selection? And these are personal positions.

What do we want to know? I say we want to know something called the response surface and utilities, and I'll elaborate on these more. But that's what we need to know.

How certain do we need to be? I say, not very.

What are we willing to assume? I would say that, given that we don't need to be very certain and the interaction between those two questions, then what we want to do is use as much as possible valid scientific knowledge of the relationship between dose, concentration or exposure, and effect.

So, let's get to what do I mean by these things. Well, a very quick course in decision theory. Decisions should maximize expected utility. I'm going to use a little notation. I'll say D sub i are the possible decisions so that decision 1, decision 2, decision 3, I'll give dose 1 or dose 2 or dose 3. I'll give drug 1 or drug 2 or drug 3. Those are decisions.

Then those are possible outcomes. The simplest case, I get blood pressure below a certain level or I don't. But there are many possible outcomes, toxicity and so on. So, each one of them is indexed and there are many of those.

Utility is the subjective value of an outcome. So, utility of an outcome is what it's worth to you or what it costs you, negative utility. What's the cost of a toxicity? What's the value of getting your blood pressure lowered? Those are subjective in the sense that they are issues for individuals to decide. If I'm going to have a side effect, it's for me to say, given that you described that side effect, what that's worth to me or what that costs me. For some people, having a limitation on their mobility might be a real tragedy. For other people, they may not feel that way. So, that's subjective, and there's not a lot we can do then as scientists about knowing about those. We have to know about utilities, but they are not something that we can do as a public activity.

Expected utility is the average of the utility across all possible outcomes, each weighted by its probability. So, there's a formula for it, but it essentially says that if you're going to get a given result that you want, and one treatment will give you a 20 percent probability of getting that result and another treatment will give you an 80 percent probability of getting that result, then the value of the first one is 20 percent times the value of the result, and the value of the second treatment is 80 percent times the value of the result. The second one obviously is better, not because the result is

better, but because it's more likely.

So, to finally get down to the bottom line of this line of argument, we said, the optimal decision, what dose you should give, what drug you should give, is whatever decision maximizes expected utility, gives you the best average payoff. The necessary empirical information, that which we can concern ourselves as scientists and regulators, that we can ask drug companies to give us or other sources are those distributions of outcomes given decisions. What's the probability of lowering your blood pressure a certain amount at a given dose? What's the probability of given side effect given that you use a certain drug? Those are the empirical things that, in the process of drug development or, in general, the process of drug use, it's designed to learn about.

So, a simple example of optimal dose is let's just say it's a binary decision, treat or not; and a binary efficacy, we get efficacy or not; and a binary toxicity, we get toxicity or not. Nothing could be simpler than that. Life is much more complicated.

I'm now even going to make it simpler and get rid of utilities. I'm going to say that the utility of toxicity is the negative utility of efficacy. I'm essentially saying whatever this side effect is, this toxicity, getting that is about as bad. In a sense it

cancels out the value of getting whatever the efficacy is. That's no so unrealistic if a drug can produce, for example, a very serious side effect and if the drug is used to treat a very serious illness -- and let's say death is the issue here -- then presumably the utility of death by either means is approximately equal and so it's not unrealistic to think that there are treatments that have this case.

But I'm not trying to make that point. I'm trying to say, let's take the very simplest case. If this were true, then it would be very simple. We would treat, if the probability of efficacy given that we treated was greater than the probability of toxicity. They've got the same utility. So, if this is more likely than that, then we treat, and if it's the other way around, we don't. It looks very simple in this very simple case.

The point I bring up here is even then here's the problem, that the probability of efficacy, given treatment, is a function of the patient and the dosage. The probability of toxicity, given treatment, is a function of the patient and the dosage, not just of the decision to treat, and this is complicated. This is what I mean then by the response surface.

That probability function, if you will, in this particular case, for example, talking about efficacy, this

would be the probability of efficacy. This is an axis that I've designed to talk about patient factors. Obviously, there are many such factors, so it's a great simplification to just have one axis on which I'm going to put all the patients.

Then there are multiple dosage decisions, as Tom pointed out. There's the length of treatment. There's the frequency of doses. There's the magnitude of each dose. I'm just summarizing them all and saying that's dose.

Then there is some kind of a surface here that describes relationship between the probability of efficacy, the kind of patient you are, and the dose you get. In this particular example, a patient down at the left-hand axis gets efficacy very quickly and it flattens out rather rapidly; whereas, a patient at the right-hand end has a little flat area where he's not getting efficacy as we raise the dose, and then he gets it and flattens out at a lower level. So, for this kind of a patient, the drug is less probable, even in its maximal effect on this individual, to cause the good effect. Well, that's one side of the equation.

Here's another side of the equation, same kind of a picture. I just made this up, a picture of toxicity which says a person at this level tends to not get toxicity

until the dose rises a fair amount; whereas, a person at this level -- maybe it's an old person or somebody like that -- starts to get toxicity earlier. Notice that this kind of a person seems to get more toxicity at the range studied here, or a higher probability of toxicity, than this kind of person ever gets.

So, we can slide this curve over to there, and then we've got the net expected utility in this very, very simple situation. That would be the maximum expected utility for somebody at this end of the scale. It's the maximum distance between that curve and this curve in this very simple case. There it is. That says that that's the optimal dose for that person.

So, the whole thing is all very nice, and the only reason I'm showing you this is not because you could ever do these computations, but because it's focusing on what do we need to know. What we need to know are those two surfaces in this simple case.

Another point to be made about this is that for this kind of a patient, there isn't any optimal dose. Toxicity always exceeds efficacy, so you don't want to use the drug in that kind of a person. So, if we could know this kind of thing, we could, in principle, settle the problem. That's what we want to know.

How certain do we want to be? I say, not very.

Why? Well, we've already heard why. Not very certain is already the current standard. Often only three or four doses are tested pre-release, and one of these is almost invariably chosen. As Carl has pointed out to you, a lot of drugs require relabeling, so it's clear we're not doing very well on this. So, we only have to do a bit better to justify what we're doing.

Also, it's true that for reasonably safe drugs, a wide dose range is tolerable, so you don't have to get it exactly right. And unpredictable inter-individual variation is always going to make it uncertain for any individual. Even if you know how to adjust the dose for age or sex or weight or renal function, there will still be unpredictable inter-individual variation, so that we'll never know the perfect dose for anybody until we treat that person.

That's the last point, which is that dose titration is a standard part of medical practice so that, in effect, we limit the harm of a wrong initial dose by adjusting it later.

So, these are all reasons why we don't have to be perfect here, but we certainly can do a lot better than we're doing and that will have a payoff not only for individuals but, of course, for society as a whole. A small increment in benefit relative to risk multiplied by

the number of people who tend to get antihypertensives can be a considerable benefit to society.

Finally, the last question is, what are we willing to assume? I say, valid scientific knowledge of dose response or dose concentration response. I guess the serious question here is why am I talking about what we're willing to assume.

Let me first set up the problem, which I have already been setting up. There's something called the curse of dimensionality and this response surface. I made a very nice little picture there. But there are obviously a large number of distinct dosage decisions, as we've already mentioned, each with multiple options. There are a large number of distinct patient variables that affect the relationship between concentration and response, and they each have multiple values.

So, here's the claim. It's impossible to study all possible combinations of dosage by patient type variables. It can't be done. Empirically it simply cannot be done.

So, response surface estimation then requires a parsimonious representation of that surface that's going to interpolate and extrapolate between and beyond what have got to be limited data. We have got to be able to describe that with less than every single type of patient given

every single size of dose and duration of therapy, et cetera.

These interpolating and extrapolating functions are the assumptions. These are the models that we use to cause the surface to be described in a parsimonious way, and if we can't make those assumptions, we cannot proceed.

The certainty assumption tradeoff then hinges on the scientific validity of those assumptions. So, to quote one of my heroes here, in other words, nothing is wrong with making assumptions. On the contrary, they're the strands that join the field of statistics to scientific disciplines. The quality of those assumptions, not their existence, is the issue. That's the certainty assumption tradeoff. That was a quote from a paper by Rob Little and Don Rubin that came out this year.

Now, how do assumptions help us? They provide the required parsimonious representation. We've talked about that already. They increase the amount of information we can recover from the data, and I'll have more to say about that in a moment. And they allow us to do what you might call meta-analysis. They allow us to synthesize information from multiple sources so that all of those various trials that Ray is talking about can be put together and all analyzed together and you can get the dose response. Ray has just made a proposal for one trial that

has several arms that are escalated in different ways, but of course, you could have different trials. There's no reason why you can't combine them, but you need models to do that. Again, I'll show you very briefly why.

Another very important part of this is that they can adjust for missing data and other deviations from protocol that make exploratory or explanatory analyses difficult because you're trying to learn what's really going on despite the fact that people didn't take the doses you asked them to or people didn't show up at the clinic for being measured when you wanted them to. So, let me just say a few words about those two so that you're clear about that.

Models increase the information recovered. Well, information is basically signal to noise, and in any given study the signal is the total variability in the data and that's kind of fixed. But the signal is the variation due to identifiable and interesting causes such as differences in dose, and the noise is the residual or unexplained or uninteresting variation that, in fact, make it more difficult to figure out what the signal is.

As I say, the total variation, which is what information is, in any given data set is fixed. So, models can't add information. They increase information by turning noise into signal. If information is the ratio of

signal to noise, if I take some noise and I move it up into the numerator, I'm getting sort of a double benefit. I'm making the noise less and the signal greater, and I'm getting more information out of this.

A simple example. The noise in a dose-response relationship due to inter-individual variation in pharmacokinetics can turn into signal regarding the concentration-response relationship if we measure concentrations and if we try to make some kind of a model that relates concentrations to effects.

Well, what's the assumption, what's the model that allows us to turn that noise, that pharmacokinetic variability, into signal about concentration response? It's simply the assumption that drugs act through concentrations, not doses. If you don't believe that, then this concentration information can never be used to get signal. If you believe that, then this is a sensible thing to do.

Turning to the other thing, models adjust for design problems. Explicit modeling of covariate effects, obviously, sex, age, et cetera, allows pooling of data from different patient groups because now we know how to adjust or we can estimate how to adjust for the different groups. But as long as we don't explicitly model the effects, then we're comparing old people to young people and we know it's

apples and oranges.

Serial observations and explicit modeling of system dynamics. In other words, making time be a covariate that's a part of our model. It shows up in our model. The time of the dose, the time of the onset of response. When we make that be a part of our model, then we can pool data that comes from studies with different treatments and different designs. And that's very important. If you observe at different times in different studies, you can't average the response at 1 week with the response at 3 weeks unless you can have somehow time in there and correcting for those features. So, you have to make time be a covariate. A model is required to do that.

Explicit modeling of variation in error structure allow pooling of data with different precision. That's a subtle point that I don't want to go into, but it's important for pooling data from various studies.

Finally, explicit modeling of deviations from protocol -- adherence and dropout being the major ones -- can avoid biases that such deviations might otherwise cause. I'll give you an illustration of that in a moment.

So, here's the illustration. Everybody who has heard me talk has heard me talk about this example, but it's a good one. Modeling in the case of Ketorolac and dose finding allows pooling of several trials. I just

mentioned that that was a feature of modeling. Allows us to avoid some bias due to deviations from protocol. In this case, it's dropout. And allows rational dose-finding which I want to end with and focus on a little more than the others.

So, the standard analgesic trial design is a parallel dose, placebo versus two to four active treatment levels. You have a pain-initiating event, for example, surgery. Analgesic dose is given when the patient says they hurt. Pain relief, which I'm going to call Y at time t, is measured on a categorical scale. Often 0 is no pain relief, 4 is complete pain relief, and 1, 2, 3 are in between.

The difficult part, the part that produces the potential bias is that if somebody is not getting enough pain relief, they can demand rescue medication, and then they're effectively dropping at some time t. That's the last scheduled observation time because after that, although of course they could be observed, the pain relief is not necessarily due to the agent you're testing but now it may be due to the rescue medication.

The Ketorolac trials. There were three of them. So, we're pooling trials here. They had different designs, but across all of them, we have these doses being tested and pain relief being elicited at these times. For

example, in one of the trials, it was at a quarter of an hour. In most of the trials, pain relief was first inquired about at either a half an hour or 1 hour. So, as I say, these are differences in designs, but we can put them together because we're going to have a model that models the time course of pain relief. So, we don't have a problem with that.

There were a total of 254 patients. As I said, our data are pain relief score and the time of rescue.

Here's just a picture of what the data look like. These are each histograms. I have six of them. This is for placebo, 2.5 milligrams, 10, 12.5, 100, and 200. Each bar is broken up into dark segments and light segments. Just look over here for an example. The dark segment down at the bottom. These are people with no pain relief, and then, for example, at this time point here, which is 1 hour, you have some people with a little bit of pain relief, more, more, and then finally dark again when you have a lot of pain relief. So, these are people with complete pain relief up here.

What we can see is as we go to higher doses, that the black up at the top tends to get more. So, the drug is working and certainly at later times. Maybe in the early times it doesn't seem to be doing very much, which is okay. There's a delay of action.

But a key point here is if you look over here, you see this placebo. What I've done is drawm the width of these bars to be proportional to the number of people who are still there at the time they're being inquired of. What you see is that in the placebo group people are dropping out. These bars are thinner than, for example, here in the 200 or the 12.5 or whatever. What's happening, of course, is the people who are not getting pain relief are dropping out. So, if you look at the people who are left, they actually have pretty decent pain relief. That's almost as large a fraction of complete pain relief there as it is for clearly efficacious dose. So, you can't just take the average. That's the problem of bias that I talked about.

Well, we have an analgesic model. There's a picture of it, and it's got really basically the following parts. It's saying that the drug dose comes in and turns into concentrations, which eventually become a concentration at some active site, and that influences the pain relief. At the same time, we've got some kind of a placebo or time evolution of pain that's causing pain, and these two are conflicting with each other. So, we get to observe the green pain relief. This model also entails the notion that there are individual effects, this gray bar here for randomness, that individuals differ with respect

to their pharmacokinetics and with respect to their pain course.

Then the model says that the degree of pain relief will influence whether or not you drop out, and I've got dark gray here to suggest that there's variability in this part too. These are probabilistic models rather than deterministic models.

What are the key assumptions that we require to analyze the data with a model like that? That drug acts only through its concentration. Notice that there's no line from dose to effect. It goes through concentrations.

Individuals are consistent over time. The idea here is that individual effects are a characteristic of individuals, but they are persistent for that individual.

Individual differences affect baseline and drug kinetics and dynamics. That's just saying that the individual effects here affecting this part, the natural time course, and they're affecting various parts of the kinetics.

The important assumption here is that this pharmacokinetic model is linear. So, if I double the dose, I double the concentration. That will allow me to extrapolate to other concentrations, but it doesn't say that the pharmacodynamic part is linear. It's not true that if I double the dose, I double the effect.

The dropout depends only on the observed level of pain relief. This is a technical point which allows me to deal with this bias. So, I have no other arrows to drop out from these individual effects, for example, or from anything else. It just says, you're going to drop out depending upon whether or not you have adequate pain relief.

So, here's the basic result. On the left-hand side, I've got the predictions and the observations of the pain relief. So, these are the data that are observed. Each of these four pictures is a kind of response surface, but it's a probability. So, this is the picture of the probability that you'll have some pain relief, that is, pain relief greater than 0, and this is that you have greater than 1, so on, all the way out to here, that pain relief is greater than 3, or in other words, this is the complete pain relief model. The z direction, the vertical direction, is the probability that you will have pain relief, as I say, here greater than nothing, some pain relief here, complete, and there in between. Along this axis I have dose and along this axis time. So, this is the response surface for dose and time against probability of pain relief.

The circles are an attempt to estimate these things directly from the data. They are simply, take the

number of people who had pain relief 2 at dose 100 and take the fraction of people who had some pain relief and plot that along that probability axis.

The surface is what's estimated by the model. But this is estimated for the observed data. So, this is the probability, for example, here of some pain relief given that you're still in the study, that your dropout time is greater than the time at which you're being observed. What you can see here is that you have generally sort of a flat curve. That's not the curve we're interested in. What we want to know here is what would happen if you didn't drop out. What is the drug actually doing, and that's estimated from the model. I can't put the data on this curve because the data don't conform to this situation of no dropout, but you can pull that out of the model.

For example, it shows you, if we just take one of the faces here, the placebo effect, which looks like it's rising in time. You saw that on the histogram. Is it really rising in time? According to the model-based analysis, it's staying flat. It's just that these are people who are dropping out. So, that's why the level is going up here. When we get rid of the dropout factor in our predictions, then we can see that there is a placebo effect, but then it flattens out.

Similarly, if you look over here, you can't really discern a dose-response curve, a difference in response at 6 hours depending upon the dose you receive, because we flattened everything out because of these dropouts. But here it's pretty clear that there is a dose-response curve and, interestingly, the dose at which we seem to get this sort of a maximum response, about 50 milligrams. So, here's a case in which the trials were done at doses well in excess of what turned out to be the maximum effect.

Now, let me just end with the dose-finding part. The optimal dose is often considered to be the minimal effective dose, and that's generally defined to be the smallest tested dose such that the null hypothesis of efficacy equal to placebo or no efficacy is rejected.

Now, Ray said we can just forget about that definition, and maybe it is a straw man, but it is certainly one that has been used a great deal in the past and I wouldn't be surprised if it continues to be used.

Well, the problem is that the minimum effective dose, defined in this way, is design dependent. It isn't something that depends on the science. It isn't something that depends directly upon the underlying situation. It depends upon the design. It depends on study size. If I make a bigger study, I can obviously find a significant

difference with a smaller dose. But why should the dose we suggest depend upon the size of the study we use to determine it? It must be one of the tested doses. You can't find some dose that you didn't test. So, again, why should the best dose be one that we happen to choose to test? It must be a single dose because that was the design of all the studies, one single dose.

But why should it be a single dose? Maybe the best way to give this drug is a dose, and then a couple of hours later, another dose, and a couple of hours later, another one.

So, this is an inappropriate use of the confirming paradigm of testing a hypothesis when we should be learning. And what is a rational approach to this?

Well, let's define these things we talked about earlier, efficacy and benefit and harm. So, let's take the outcome that we want to maintain a certain minimum probability of pain relief greater than some value. So, K is some value. That could be 3, in which case we want to say we want to maintain a minimum probability we have complete pain relief. But if K is 0, which is the example I'll show you, then we're saying we want to maintain a certain minimum probability that we have some pain relief between certain times and we're going to put a large utility on that.

Our "toxicity," since we didn't have any toxicity in this study is just dose magnitude. What we're going to say is the bigger the dose, the more it costs us. We're not going to big disutility on that, so that the combination of these two things will essentially yield us the smallest dose that will produce the minimum relief that we want between those times.

Patient factors. There won't be any. No difference between men, women, or whatever. But we could have put them in there.

The response surface that we're going to use is the one that I showed you a picture of. It's the efficacy outcome obtained by a simulation from the model of the probability of pain relief being greater at any given level, and we have that model because we fit it to our data.

To make things clear that we don't necessarily have to use that minimum effective dose kind of idea, we say, let's find the two-dose regimen where we give one dose at time 0 and we give another dose at some other time. What we want to find is that other time and the value of that dose plus the value of the time 0 dose. So, there are three things we need to find out. Two dose sizes and the time of the second dose, something we never studied. And we want to find it such that the probability that we have

some pain relief at all times between 1.5 and 8 hours is no less than some minimum.

Before I get to that minimum, let me just talk about the 1.5 to 8 hours. There's a time lag in the effect, so we don't see any pain relief before an hour. So, it doesn't make sense to make that earlier, and I chose 8 hours just to go beyond the data. The data in this study only went to 6 hours. I just want to show extrapolation.

Now, what's the level of pain relief? I want to have a greater than 70 percent probability that I have some pain relief. Why? Because that's the value that you get between that time from the model for a single dose of 200 milligrams given at time 0. That was the largest dose given. So, I want to match the efficacy of a single dose given at 200 by giving two doses at different times.

Here's kind of a picture of the initial dose versus the second dose. What this says is if I give 60 milligrams to start with and 60 milligrams at 6 hours later, I'll get just a good a result as if I get 200 at time 0 over 1.5 to 8 hours. If I give about 50 milligrams at my initial dose or 40 and then 30 or 20 at 5 hours later, I get the same result. And if I give 20 at time 0 and 40 at 3 hours later, then I'm giving a total of 60 milligrams, and I'm getting just as good a result as if I give 200 at time 0. The winner actually is 4 hours. If I

give 20 at time 0 and 20 at 4 hours, I get just as good a result at a total dose of 40 milligrams versus 200 milligrams if I use it that way.

Now, if you think it's better to give 40 than 200, here's an example of a rational approach to dosage. You don't have to use these same criteria, but this is the idea. What do you want to know? What are you trying to achieve? Now, let's use the model to tell us how we get it.

So, optimal dosage decisions maximize expected utility. Utilities are subjective values of outcomes. Expected utility is an average over outcomes, each weighted by the probability. And the set of probabilities is the response surface. It's a function of dosage regimen and patient features. It's derived through experiment and observation.

I have not shown you, as Ray promised that I would, how you can analyze an escalating dose trial and find out about those models. I showed you an arrow dose trial. But again, that's a technical issue. It can be done. You use these things called mixed-effects models, and you can do it.

Response surface estimation is best viewed as a learning, not a confirming exercise. You can't study all the possible combinations of regimens and patient types, so

we have to interpolate and extrapolate. I showed you interpolating and extrapolating as an example of doses for the Ketorolac. We went beyond the time of the experiment and we used two doses rather than one.

PK/PD model-based learning designs analysis can do so by making scientifically valid assumptions. Assumptions reduce inferential certainty because if the assumptions are wrong, the conclusions are wrong, but fortunately dose-finding doesn't require great certainty.

Thank you.

DR. PACKER: We'll take questions from the committee. Any points of clarification specifically focused on this presentation or any residual questions to Ray's supplemental presentation? Tom?

DR. FLEMING: I've got a number of thoughts. I think much of what Lew has said I find myself in agreement with. I would maybe refine some of the perspectives, at least as I think about them.

Just as a quick aside, even this isn't the main focus, Lew had pointed out the importance of three questions -- what do we want to know, how certain do we need to be, and what are we willing to assume -- and pointed out the distinction between domain-specific experts and technical experts and more or less indicated that the domain-specific experts are the key people for answering

those three questions, and once those answers are in, the technical experts are then charged with essentially the scientific modeling and analysis of data.

In a real sense, I would agree with how I would allocate primary responsibilities, but I would argue that the domain-specific experts need to have at least a given level of understanding of what the technical experts are about so that they have at least a general sense of the reliability of the analyses by the technical experts and how much the model assumptions can impact on that reliability.

Conversely, I would say the technical experts really need to have a base understanding of the clinical situation since they're the ones who really understand the robustness of the results from the modeling, they have a general sense of what makes reasonable justification in terms of the answers to those questions. In fact, I've often heard it said that in an ideal situation if you sat down at a meeting of the technical experts and the domain-specific experts addressing issues such as this, you ought not be able to tell initially which is which because both disciplines should have a given understanding in order for this synergy to really achieve its maximal.

Relative to the --

DR. SHEINER: Can I just make a quick response

to that? Because I think we are pretty much in line.

There are two best of all possible worlds. One of them is the last one you described, which is that you can't tell the difference. I don't know that that's achievable. The other best possible world, of course, is that everybody is absolutely clear on just where their expertise is and makes completely clear to their counterpart just what their assumptions are.

Because that's also impossible, that's why you have to have what you were talking about, Tom. That is to say, you need to be in the same room together because there are always hidden assumptions that we just kind of go on and don't even know we're making that the domain expert will make or the statistician will make. Each one has to be able to say to the other one, how come we got there from there? What did it take to do that? But as I say, there are these two alternative best possible worlds, and if we could be in either one, that would be fine.

DR. FLEMING: To move to the other aspects of the presentation, it is exactly as you point out, Lew, critical to be able to augment the actual data with modeling in order to try to glean as much insight as possible from the data.

I would agree very much with what I think I heard you say toward the end, and that is the amount of

reliance that we're able to make on modeling depends on whether we view ourselves to be in an exploratory versus a confirmatory stage of the overall development. I see, exactly as you pointed out, the IIb stage as a critically important stage for exploration and, in turn, an invaluable contribution to that is modeling.

Of course, not all modeling is the same. Certain models carry with them stronger assumptions and certain types of modeling assumptions are stronger. It's important to distinguish those models that are making very strong assumptions that in fact in a certain sense can give us broader insights, but those insights may be at some risk in terms of their reliability.

So, to be more specific, if we're talking about modeling covariate effects or modeling for dose response, I think many of us are more comfortable if our conclusions are based on interpolations as opposed to extrapolations. So, for example, if we're looking at the effect by age and we're really interested in the effect in 70-year-olds and our data is only on effects in 20- and 30-year-olds, then that extrapolation puts us at much greater concern than if the data is looking at results in 20-year-olds and 70-year-olds and we're interested in 50-year-olds. We're interpolating rather than extrapolating.

And same issue occurs in dose response. If

we're trying to understand this dose-response surface and we fit a model and we only have information on lower doses, I'm very reluctant to extrapolate to a high dose, whereas I'm much more willing to interpolate.

And this same issue arises for looking at models over time as well.

This issue is actually something that occurs across a wide range of challenges that we face. Do we need to randomize? Well, technically we can adjust for differences between intervention and control groups based on baseline covariates that we would have. Unfortunately, those baseline covariates are just the tip of the iceberg for what explains the differences, and hence, that's the concept of extrapolation. That's the reason we argue we have to randomize because the covariates that we have for adjustment are just a fraction of the totality of covariates that we would need.

Is missing data important? Is it informative missingness? Well, if we can explain the mechanism of missingness by covariates, then we can fully adjust for the potential bias. But it's the same situation as randomization. We only know a tip of the iceberg of those covariates that truly explain the nature of missingness in most situations.

Surrogate endpoints. Why is it that we often

don't accept a surrogate endpoint? Well, if we understood the causal pathways by which the disease process influences the outcome and those causal pathways are entirely mediated through the surrogate marker of interest and there are no unintended effects of the intervention, then we can rely on the surrogate. But again, it's an extrapolation problem. We're concerned that the effects of the disease process are only partially mediated through the biologic marker of interest, and there could be unexplained covariates.

So, basically what I'm trying to get at is the fundamental scientific principle here that very often we have abilities to set up models and those models are models we feel much more comfortable trusting if we are, in essence, interpolating as opposed to being forced to extrapolate. It's the reason I have trouble with non-randomized trials or with missing data or with reliance on surrogate endpoints, or in the case of discussion here, fitting dose-response surfaces or looking at effects over time or looking at effects by covariates. If I have data and I can interpolate, I'm going to be more comfortable than if I have to extrapolate.

But my level of comfort is much greater is you're asking me to do this in an exploratory approach. So, what you've laid out makes perfect sense for me, especially in a phase IIb trial where I can, in fact,

arrive at the strongest possible conclusions from the data that you have given to me, the best insights possible, but then I can come back and confirm that in a confirmatory trial.

DR. SHEINER: All the principles that you enunciate are ones that I agree with, but I think there is a disagreement and I would like to sharpen it.

I think the basis on which you trust making a prediction of an unobserved result, interpolation or extrapolation, is not whether you're interpolating or extrapolating. The basis is the credibility of the assumptions required to make that prediction.

Now, interpolation requires, especially if it's narrow interpolation, only the assumption of smoothness. It doesn't require any knowledge of process. Almost everything is smooth.

DR. FLEMING: Or at least monotonicity.

DR. SHEINER: Yes, monotonicity or something like that.

Whereas extrapolation, especially distant extrapolation, requires that the basis on which that extrapolation is made have actually credibility in terms of -- you're remodeling the real world -- the real world. So, I'd say it's how good the science is, and the fact that you can rely on smoothness or monotonicity in many fields and

you don't have to think too much about it does definitely make you feel better.

But I think the only solution to this problem, which is going as far as you can go with information, giving people the best information about how to dose patients, the only way you're going to get there is by carefully looking at which scientific assumptions you need to make, making sure you've got those straight, and then letting that make you be comfortable with your extrapolations.

So, again, the distinction between the exploratory and confirmatory for me in this particular instance is one in which we cannot possibly confirm every possible combination of dose and patient. So, we are going to have to be left, in the end, making recommendations based on exploratory or explanatory statements. Therefore, our obligation is to make those as good as possible within the constraints of how much money we're willing to spend, how much time we're willing to spend, and that means getting the science straight. That's the key thing.

In the end, the label will be written with respect to dosage based on models or it will be written based on empirical observations, in which case it will be totally inadequate because it will be a dose you tested which you didn't know whether it was the right dose, and it

will be a pattern of testing which you didn't know was the right one. So, we've got that tradeoff.

We want to be certain that we know what the dose does. We have to suggest one of the doses we tested. On the other hand, there's no reason to believe, as Ray has shown us very nicely this morning, that the doses we tested are going to be the best doses.

So, I claim that, no, we can do better than that. We can do reasonable interpolation and extrapolation and we will have to argue about that. We will have to reach a compromise, and it will be the domain experts ultimately who will say I trust that assumption. You can't block the beta system any more than completely. Consequently, this thing is flat after a while, whereas here's a drug that operates by unknown mechanism and I'm not exactly sure that it's going to be flat within some reasonable range. All kinds of things like that.

So, I think that's a difference in that I'm saying we are really going to have base the best possible recommendations on assumptions that we have not demonstrated in this case in detail.

DR. PACKER: I just want to clarify. Tom, what I did not hear Lew do was advocate a departure from randomized trials which would address many of the assumptions that you're most uncomfortable with in terms of

accounting for unknown baseline covariates. I didn't hear, Lew, that you advocated such a departure.

DR. SHEINER: No. Again, in the general case, I'll advocate using totally observational, uncontrolled studies if that's the best you've got and that's all you can do. I always want to learn the most I can because I've got to make a decision. You have to make a decision. And to base a decision on less than everything you know is a mistake.

But randomization. Absolutely crucial. Controls. Absolutely crucial. Nobody is suggesting that we abandon those. Nobody.

DR. PACKER: My sense is that the preservation of randomization minimizes a number of assumptions you make that are of low quality.

DR. SHEINER: I think Don will talk about this more. The more you have departure from protocol -- so, let's say people don't take the doses that you prescribe and yet you still want to know what happens when they do take the dose -- then you have to use those actual departures -- and they break the randomization somewhat because you didn't randomize for people who were going to not comply. They just did it or didn't. So, if you strictly hold to the randomization, then you get certain nice features, but you answer what I consider to be the

wrong question. The use effect in this question rather than a method effect in this question. What happens when you prescribe the drug rather than what happens when you take the drug.

On the other hand, if you don't start with randomization, you don't have a chance, and if you can design protocols that patients accept so that they follow the protocols, then you've got the benefits of both. Then you can do a nice explanatory analysis, if you design it well, and you've got the benefits of the causal inference that Donald talked about more that's based on randomization. But serious causal inference has to be based on randomization if you've got any doubt at all about the underlying assumptions.

DR. PACKER: Carl?

DR. PECK: Well, I had wanted to comment on Lew's use of the word "assumptions," and I'll do that but Tom's comments actually allow me to say this in the context of his.

What I want to say was contemporary drug development programs, ones that are being run right now and the ones with clinical trials appearing in NDAs, almost all of the trials are randomized. So, the issue of randomization is really not an issue except in oncology. Oncology is sort of in a class of its own, not employing

randomized trials except perhaps in a single phase III, and that's a special problem.

But when we count the 30 to 50 to 80 clinical trials in programs we're looking at, three-quarters or sometimes 90 percent or sometimes every single trial has had a randomized phase in it. They're almost all blinded. So, the quality of data coming from these clinical trials is very high from the point of view of randomization and blinding.

So, the point I wanted to make is the "assumptions" that can be available at the time of the IIb trial, exploratory dose-response trial -- typically there will be 30 to 50 trials that have preceded that trial. There's a huge database from randomized, blinded trials, surrogate endpoints rarely, biomarkers that have putative linkage in a mechanistic way, often definitely clinical adverse reaction data that, if properly analyzed and integrated, provide a very strong set of "assumptions" that I would -- I don't like the word "assumptions" at this point because I think that they're so data-informed that they reflect a reality that's more than just an assumption pulled out of the air.

But I don't think we need to worry about the lack of randomization. It's the rare clinical trial today that's not randomized.

DR. FLEMING: Just to come back to my earlier comments, I'm using the concept of randomization to simply illustrate the point of where many of us have drawn a line in terms of our willingness to extrapolate. So, specifically the focus of the discussion, as I see it today, is on the model assumptions associated with dose-response surfaces, as well covariate effects and how they influence those dose-response surfaces.

But I wanted to point out that we obviously in our research strategies make model assumptions in many other ways or are unwilling to make model assumptions in other ways. I used as examples of those other ways randomization, handling of missing information, and surrogate endpoints. My comment is, in those three specific instances, the reason many of us believe in randomization and in the need for minimizing missingness of information and using clinical endpoints is the alternative of using non-randomized trials and models for missingness and surrogate endpoints requires very significant model assumptions that are more in the spirit of extrapolation. I.e., we're having to make assumptions beyond the covariates we have in a model to specifically model the necessary relationships.

So, using that as a guide, I was arguing that, in the context of our discussion today, which is mostly

focusing on dose-response surfaces and covariate effects, I'm much more confident with making conclusions about given safety and efficacy at doses if I'm interpolating as opposed to having to extrapolate. Every single model that Ray put up at the beginning of today assumes monotonicity overall. It's not necessarily the case that efficacy is monotonic. I'm more comfortable, though, with an assumption of monotonicity when I'm interpolating than when I'm having to extrapolate.

DR. PECK: Well, my comment about interpolation and extrapolation and the interpretation of phase IIb data is this is sort of an accident of the moment where we have incompletely analyzed the available information. You cited, for example, you'd be uneasy about extrapolating beyond the maximum age that had been studied, say, on a phase IIb trial. Many of us would be bolder than that because we know enough generally from these trials about what the mechanistic elements of the aging process are that are affecting drug disposition and pharmacodynamics.

So, if your maximum age is 50 and you're wondering what happens in the 80-year-old, you can use data-informed modeling to understand that between 50 and 80, your creatinine clearance is going to naturally decline another 20 to 30 percent, that the drug is excreted primarily in the urine, and that there's a dose and

concentration effect related to side effect profile. You can certainly expect a higher side effect profile in the 80-year-old given a common dose and the expected higher blood levels. So, I would not feel my hands tied at extrapolation in phase IIb trials, given that the data that preceded it is adequately incorporated. You see, it's not a frequentist situation at the end of phase II. You are standing on a mountain of data and potential knowledge if you extract it.

DR. PACKER: I don't want to speak for Tom, but I think what he's saying is that he's not as convinced that we are that smart. I think Lew's point specific to this is the fact that the extrapolation requires assumptions from domain-specific experts who aren't that smart either. In fact, there may be other determinants of how an elderly patient responds to a treatment other than the fact that there is renal impairment in the elderly and one can measure blood levels in renally impaired as a function of creatinine clearance. There are millions of changes that occur with aging that can affect response and renal excretion is only one of them.

My sense is that Tom is a little bit concerned that an extrapolation based on a single biological determinant just isn't good enough to allow for this, and if domain-specific experts say that that's the primary

determinant, then the answer you may get may, in fact, be the wrong answer. I hope I said that correctly.

DR. SHEINER: This is such a nice example, this whole thing, what we're doing now because it just is so clarifying, the difference between confirming and learning. It's clarifying the difference between the attitude you take when there is no cost to being right or wrong, so to speak. That is to say, you're just trying to figure out how the world works. And science I think appropriately protects scientists from having to pay a big penalty if they make a wrong guess. Well, that's it. You still get paid your salary and somebody corrects it and you keep on going.

But we're now talking about making suggestions. So, what's the competition here? The competition is either saying nothing about the effects of age, in which case you are leaving people who are less domain expert than the domain experts that you're able to talk to to fend for themselves, or demanding that age be studied. And then somebody else will come along and say, well, sex has to be studied and now this has to be studied. And there's got to be a stop somewhere.

Either you study everything a little bit with a few people, in which case you've got to be very Bayesian about how you interpret those data because those are very

few points, so you have to put them in a very strong assumption-rich context in order to do that. Maybe you want to do that. Or you put a huge effort into studying a bunch of old people and a bunch of young people, and then you don't study a bunch of women and a bunch of men.

So, you've got no choice. You've got to make a recommendation. It seems to me that backing off and saying, well, a million things might happen here and we don't know them all is actually doing a disservice to the public because you know -- you being the manufacturer, the experts, et cetera -- as much as there is to be known, and it's your obligation to communicate that in as clear a way, with the right hedges, as possible. That's your obligation.

Now, it's quite different than your obligation as a scientist if you're just trying to find out knowledge. Then you say, look, I didn't nail this one down. So, this hypothesis remains open. I encourage other people to investigate it.

DR. FLEMING: It certainly is exactly as you say, Lew. The far and away most preferred strategy is what I call to be inclusive when I'm looking at eligibility requirements in clinical trials and, as has been advocated here, inclusive when you're looking at ranges of plausible doses or schedules or schemes that could achieve optimal

benefit to risk.

I think sponsors are well served by trying to achieve that inclusiveness. Those insights can readily in many settings lead to a better choice of dose and schedule that would optimize benefit and reduce risk. By looking at being inclusive in eligibility criteria, it allows us to gain insights about whether, even if we can't say that they are definitive insights about subgroups, at least it gives us the opportunity in phase III trials to generate clues because, as you point out, ultimately we're going to have to make decisions as caregivers as to whether to use an intervention and if so, in what way.

From a regulatory perspective, I would say that considerations are also very consistent as they would be from the perspective of the clinical community and the insights that you'd want to know. Ultimately, though, we want to be able to have evidence that we view to be adequately convincing or conclusive that a given regimen, as it's being proposed to be delivered, is in fact efficacious and has a safety profile that's consistent with a favorable benefit-to-risk profile.

DR. SHEINER: Absolutely. I'm not advocating abandoning having to have some empirical evidence that something you did worked out in the world in fact as observed better than if you hadn't done it or the control.

I absolutely agree you need confirmation. It's just the next thing is what do you put that within. I don't think we disagree.

I think actually a continuing dialogue on this, we'll see, in the end comes down to that slide that I took from Don. It has to do with the quality of the assumptions and that interaction between people who understand the implications of assumptions and the people who know the field well enough to know how much credibility to put on them.

DR. PACKER: I think that this would be a very good time to take a lunch break. Let us take a break for about, say, 35-40 minutes. Make it 35.

DR. LIPICKY: We can't do it in 35.

DR. PACKER: We'll reconvene at 1 o'clock, and at 1 o'clock we'll start with Dr. Rubin's presentation and move forward.

(Whereupon, at 12:21 p.m., the committee was recessed, to reconvene at 1:00 p.m., this same day.)

AFTERNOON SESSION

(1:15 p.m.)

DR. PACKER: We're going to resume by asking Don Rubin to cover the topic "How Can We Know What We Want to Know?" Don, why don't you proceed with your presentation.

DR. RUBIN: I don't have any slides or transparencies to show you. I'm going to be just speaking.

We began with Carl Peck discussing whether even the analyses of current data that are collected are adequate for the purpose of decisions about dose response. Carl basically came to the conclusion that the current analyses were not. Even though the trials are often designed as dose-response trials, they're analyzed in this traditional, null hypothesis testing dichotomy which did not yield very helpful information for many decisions.

Then Ray extended that to say maybe the data that are collected are inadequate, inadequate in the range of doses, and suggested that we really should be doing trials with a greater variety of dose ranges than we have now.

I tend to agree with both those positions.

Lew has added support for those positions, showing you explicit examples and discussing how you might analyze the data.

I'm just going to make some general kind of comments from a statistical point of view. I'm relatively recent into this area, but I have gotten involved in clinical trials and the complications in clinical trials recently, complications such as missing data, noncompliance, and how to analyze trials in that context. I guess at the outset I should say that I completely support the idea that randomization should be done. We're not talking about dropping randomized trials at all. But I think we will be talking about trying to do a better job of analyzing the data from randomized trials and more flexible randomized trials than we have now.

One thing that I want to mention is that there's a context that I think is fairly interesting about the causal inference randomized trials beyond drug development. The context I'm thinking of right now is in social science. A couple weeks ago, I was at a conference dealing with social science, and one of the topics that was discussed was the tremendous advances that have been made in social science in thinking about causal inference. In that context, David Cox was there and I had a chance to talk to him about something. I think it's kind of interesting.

When do you get to use the word "causes" in dealing with data? He had this example. He said, let's

suppose you have a huge, randomized experiment with no complications at all. There are no complications with missing data, dropout, noncompliance, nothing like that. And there's an obvious effect shown. Do you get to say that the treatment causes the outcome?

Well, the interesting thing was that David said that he really wanted to see some evidence of mechanism before asserting that the treatment caused the outcome. Now, why is that?

I'm not sure when you want to use the word "cause" or not. I'd be inclined to use the word "cause" and just say I don't understand the causal mechanism.

But David's point was that when you're dealing with a trial in a particular context, what you're really interested in is in generalizing outside that context. Unless you have some understanding of the mechanism involved, you can't really assert how well it might generalize.

That's a comment, I think, on intention-to-treat analysis as well.

Why do I say that? Well, intention-to-treat analysis is very much viewed as a confirmatory kind of analysis. You confirm the results of it, but how well do you know that the results that you see can generalize? If you think about the situation with noncompliance, for

example, when you have noncompliance in a trial, as we all know, the standard way to deal with it is an intention-to-treat analysis. The problem with intention-to-treat analysis is it bundles together personal behavior and compliance behavior in this trial with a scientific effect of the drug.

Now, I'm not arguing for as-treated. I'm certainly not arguing for that kind of analysis, nor am I arguing for per-protocol, nor am I arguing for anything simple like that. But I am arguing for trying to disentangle the effects of behavior, which probably will not generalize outside this trial, to the effects of the drug, of actually taking the drug, which probably will generalize outside the trial. And I've written some recent papers on that, which I hope have had some influence.

The thing that you expect to generalize beyond a trial is the science. The scientific part is in small steps. That's an important idea, that the small steps, the small things are the things that you have a better understanding of.

The words "confirmatory" are a little ambiguous. I think they promise more than they really deliver. If we are in a perfect randomized trial, why do these results apply to people in the future? Do we really have a random sample of people in the future? No. Why do

we think they'll generalize then? Well, because we have some confidence and smoothness of the world, that the effect of the drug is not going to be affected by ozone layers or ozone counts or something else like that. So, we really never have the ideal randomized experiments because randomized experiments would take units from the future. So, we're always relying for the generalization on some kind of science.

Now, the information that we require for approval of a drug is a dichotomous decision. You eventually have to approve it or not approve it. That information is very different from the information that's used by docs for making decisions because those decisions are not dichotomous decisions. They're not up or down decisions. They're not yes or no decisions.

One of the principles of statistics is that you should avoid summarizing data, avoid making summaries along the way until you really must, until you're really forced to do so. That's part of the decision theory, the analysis that Lew has put up, that you want to weight things by probabilities of their occurrence and you want to wait till the very end until you make a decision.

The issue is that as far as the use of dose-response trials by docs, they never really do have to make a dichotomous decision until the time when they make a

recommendation for a patient.

So, what is a real ideal labeling for a drug? Lew has pointed out the response surface. I think what we really would like to do, if we could, would be to have a real response surface that would have something like the following. We have all these covariates that describe the patient and the situation. At each value of these covariates, you'd like to have the expected value of the effect, the variability of the effect for patients of that type. You'd like to have the expected number of side effects, the type of side effects, and you'd like to have standard errors, measures of uncertainty about all those things.

Ideally, that's the information that you'd supply with an approved drug, not that the decision to approve or not to approve has to be continuous, but the information that you provide will be this continuous kind of information. That information is very, very detailed. I don't know whether we can ever move towards that, but that's the kind of information that would be most useful in actual prescribing use. It's really this ideal of individualized medicine, very personalized medicine. Ideally that's what a label would be in some generalized sense of what a label is.

I'm sort of cycling around here. I hope I'm

not cycling around too much.

What's the purpose of a trial? Is it just to test the null hypothesis or is it to summarize evidence? Now, I think many of us would think that ideally you would like to summarize evidence in the trial about the efficacy of a drug, and that means the whole dose-response curve. That's something that I think we heard pretty much agreement on, that that would be the ideal thing to do. That's sort of a Bayesian perspective. It's the whole modeling perspective. You want to summarize what we know about the drug and not test the null hypothesis.

When you think about testing a null hypothesis, it's actually not a very interesting thing to do. If you get a very significant p value, what does it really mean? Well, what it really means fundamentally is I have found a model that doesn't fit, and it really doesn't fit. And the more significant, the more convinced you are it doesn't fit. And that's not very exciting.

Why are we so focused on that? We're focused on it because usually we have an alternative model in mind that we're willing to accept in place of the null. But you have to admit it's not very interesting to just say I've found something that doesn't work that is a significant result. I found a model that doesn't work.

So, why do we do that? Well, it's a pretty

easy question to answer, and I think the tradition of statistics, as it's evolved over the last half century, has been focusing on questions that are easy to answer and not answering the right questions. I think John Tukey once said that it's better to address the right question, even if it's a poor answer, than to address the wrong question and get a good answer for it. I think what statistics as a professional field has been too oriented at doing is getting a precise answer to the wrong question, or often the wrong question, the precise answer to a too simplistic question.

Now, if you address the more general questions, questions that are of interest, you often do get somewhat non-robust answers to the real question. Now, this is a point that Tom Fleming made. There are different questions. If they're more like interpolation, you certainly get more robust answers than if they're more like extrapolation. I think the difference between interpolation and extrapolation is kind of difficult to address in many cases. In some cases it's pretty easy. In the case that Ray was describing where he showed plots, some of those plots where you were extrapolating to beyond a dose that you had ever seen. That's probably like extrapolation, although Carl had this example of age where there's other data that you have available that would

suggest you really pretty much know what's going on there.

But I do think if you take the attitude that the right thing to do is to try to get an answer, even if fairly non-robust, to the right question, that you actually do better in real practice with that.

So, what am I saying? Well, if we accept the objective as to try to get this response surface which is a multivariate response surface, it's the expected value, the variance of the effect, the side effects, as a function of all these covariates and dosing levels, that's a hard question to address and you'll have to make a variety of assumptions when addressing it. But it is most related to practice. It is a question that really does address how practice works.

If you do get a non-robust answer to that question, what you often will find is if you're looking to that formulation, that response surface to address a simpler question, an average of some kind, then when you average over that function, you will get robust answers. The margins of a high dimensional surface are often very well estimated and very robust even if you're using a model to do that. So, what I'm saying is that if you use a model to try to address the right question, that you'll often get a great deal of uncertainty in the answer to that right question, but then when you integrate over that model to

ask simpler questions, you'll get robust answers to those simpler questions. The averages are robust.

But for actual clinical practice, you'll get no worse decisions. You'll get better decisions, in fact, than if you ignored the right question and provided no evidence at all, no argument at all for what the right question should be because then you're left -- I think this point came up at the end of maybe Lewis' talk. If you just do not provide information, then the user, the doc has to make decisions based on no information at all. If you can provide some information, it's better than none.

Now, I want to say a couple of other comments. Then I'm going to close.

I did say that no one is saying don't do randomized trials. Absolutely, we have to do randomized trials. In design, you must do it.

However, how do you deal with missing data in a randomized trial? Well, you really can't if you're going to try to do it in a completely robust way and try to put in worst values or something like that, but that's not getting at science at all. When dealing with noncompliance, a similar kind of problem. When you deal with surrogate outcomes, you have a similar problem.

Where I started out with the David Cox comment was the fact that things that generalize best are small

pieces of scientific understanding. In order to decompose those, even in a randomized trial, you're going to have to rely on models to do that. I think surrogate outcomes is an area which is very poorly understood. I think future technical statistical work will have a big effect there in understanding how to do that. That's in parallel with understanding, dealing with noncompliance in clinical trials, how to decompose the effect into the behavioral effect on compliance and the effect of the drug.

I think that's all the comments I really want to make, and I will be happy to take any questions. I think that there should be some.

DR. PACKER: Any questions from anyone on the committee? Actually we can open up for any of the presenters today. Ray?

DR. LIPICKY: I'm just curious to know how you define scientific.

DR. RUBIN: Well, I'd say that's obviously a tricky question. Scientific really means there are relationships among variables that we think we understand in kind of a mechanistic way, not necessarily really a deterministic way, but a mechanistic way. Things such as reasons for dropping out of a study are not part of science.

DR. LIPICKY: Right, I understand.

DR. RUBIN: They're reasons for behavior, not part of science.

DR. LIPICKY: So, let's say that I have three sets of data that are really tables of random numbers, but I can deduce from each set something that the data says to me and I can somehow or another relate those three data sets to one another. Do I have a scientific piece of stuff?

DR. RUBIN: I'm not sure I understood the question.

DR. LIPICKY: I have three tables of random numbers. I've done an experiment, and I have three tables of random numbers. I look at the first table and I say, I have dose and effect here and I can make something out of that. But really the relationship between dose and effect is --

DR. RUBIN: Noise.

DR. LIPICKY: -- just noise. But I can make something out of it. So, now I think I have found something.

DR. RUBIN: And you can make something out of it because?

DR. LIPICKY: No. I can. I could. Believe me, I can.

(Laughter.)

DR. RUBIN: If you keep working at it, you can. Is that what you're saying?

(Laughter.)

DR. LIPICKY: Then I have another table of random numbers, and I do some similar, clever thing, analysis. Then I have a third table of random numbers, and I do the same thing there. So, now I have three different pieces of relationships. Would you call that scientific?

DR. RUBIN: It sure doesn't sound it to me. I'm not sure what kind of analysis you're doing.

DR. LIPICKY: Well, I guess really let me ask the question differently then. The question I'm asking is in this set of data that you wish to analyze somehow, don't you have to have something in that data set that allows you to say that this is not a table of random numbers?

DR. RUBIN: Sure.

DR. LIPICKY: Yes, and what is that something that you have that allows you to decide that? Because I can fit some kind of a function to anything, and I'll call that my model.

DR. RUBIN: I think that you're confusing a couple different things.

DR. LIPICKY: Okay. That's what I'm asking you to straighten out.

DR. RUBIN: Right, and I'll try to straighten

it out.

One is a paradigm that says before you begin an analysis, before you look at the data, you should have a specified plan of attack. Another thing is that that plan of attack should end with a p value, and those are two very different things.

DR. LIPICKY: How do I get a p value out of an Emax model?

DR. RUBIN: If you're going to fit an Emax model, how would you get a statistic out of it?

DR. LIPICKY: How do I get a p value.

DR. RUBIN: A p value. You do a randomization test.

DR. LIPICKY: What is the test?

DR. RUBIN: The test is you take whatever parameter you fit from the Emax model and you run it through the randomization distribution. Then you get a p value for it. That's completely valid.

DR. LIPICKY: You'll have to be more explicit. The last math course I had was in high school.

DR. RUBIN: Something that's often confused in statistics is the value of randomization and what it brings to you and the traditional analysis that you do in a randomize experiment. There's no reason why the only statistic that's of interest is a simple comparison between

two means. Randomization doesn't have anything to say but it justifies that.

If you look at the original work on -- where does randomization starts? It starts with 1925, 1923 with Fisher and Neyman. What was the great innovation about it? Well, everybody was sort of talking about randomized trials before then. They were writing papers saying, if you assigned the plots as if they had been random....

Fisher's great innovation in 1925 was he actually proposed to physically randomize the plots. He said that physical randomization would provide an internal test of the effect.

Actually Neyman did a paper in 1923, two years before Fisher proposed that, doing mathematical analysis with a randomized experiment. And mathematical analyses of randomized experiments as if they had been physically randomized go back five years before that. That sort of spirit was in the air in the 1920s.

But Fisher's great innovation was he actually said do it, do a randomized experiment, randomly assign the units. Why? Well, if you randomly assign the units, then you will have all permutations. You'll have laid out in your mind at least or on a piece of paper all possible permutations of the units that would be done if you had done randomization. So, if you have 20 units, 20 people,

and 10 of them are randomly assigned, one treatment, and 10 of them are randomly assigned to another treatment, then there are 20, choose 10, randomizations.

What is a p value? A p value is like a proof by contradiction. How do you prove something in mathematics by contradiction? How do you prove that the square root of 2 is irrational? What do you do? You start off by assuming that it's the ratio of two relatively prime integers, and then you reach a conclusion that they're both even, that these two integers are both even. I've reached a conclusion. Therefore, what I assumed must be wrong.

So, what is a randomization? What is a p value? A p value is what do you want to prove is wrong? What you want to prove is wrong is both treatments have the identical effect. Absolutely everywhere it's an identical effect. So, if you have two treatments and you have these 20 people and I observe a value under one treatment for each person, but not the same treatment obviously, under the null hypothesis, you know the value under the other treatment. So, you know all the values that you might have ever observed no matter what the randomization would be. If all possible randomizations, you can calculate the value of the statistic.

DR. PACKER: Ray?

DR. LIPICKY: Okay.

DR. RUBIN: I'm almost done, if I may complete this.

DR. LIPICKY: I don't understand what he's talking about.

DR. PACKER: I think that's okay.

(Laughter.)

DR. RUBIN: The reason for doing this is just trying to say what a p value really is. I don't see how you can be in love in a p value once you understand what it is. I'm just saying that no matter what statistic you've put through this randomization test, it's completely valid and it has the same interpretation as a standard test for the means. You've got no obligation to run the means through. You can put any test of this that you want to, and it could be your estimate of an Emax parameter that you could put through, and it's just as valid as a test of the means, as long as it's the one you decided a priori, that's the one you're going to put through it.

DR. PACKER: Really, Lew? Do you have a question?

DR. SHEINER: I want to bring it to the context that you're dealing with here. One of the things that advisory committees ask about is the science, which is sort of what Don is talking about. One of the things you ask about is what's the import in the world.

One thing that you haven't addressed, Don, and I think maybe we could all think about a little bit, is that there's an extra, little hooker in the whole thing, which has got to do with the fact that, not maligning anybody's intentions, the temptation to make money or to be an enthusiast for something, in this case profit, is more than the best of us can resist. Consequently, there's the extra part of what does the FDA require to be sure that nobody is playing the game not straight, which is a little different than our usual assumption, which is that whatever data you show me, of course, they're the right data. You may have interpreted them wrong, but that's them. I'm not even talking about cheating here.

I'm just saying one of the problems that has been raised about model-based stuff -- and I'm going to be the devil's advocate here -- is that if you get to decide what you want to do with that, as Ray I think was suggesting, you can always do something that will make it look a little better for you. So, society has the extra problem of perhaps costing itself a little bit in terms of the ultimate amount of knowledge they get out of data in order to protect itself against the temptation. And the question is, how do you build that in?

DR. PACKER: Let me ask a question because I think it's essential for us to bring the concepts that have

been elucidated and focus them on the topic today. Lew, let me ask you a question directly related to what you have just brought up. If sponsors are going to go forward and elucidate the response surface, to what degree do the models that they would use need to be prespecified?

DR. SHEINER: That's the key issue. Exactly what does prespecification mean? Does it mean prespecifying the model, or does it mean prespecifying a procedure for coming up with a model?

Let's just say I'll do a multiple regression, and I'll pick the variables that meet a certain criterion. Then I'll test the predictive ability through some kind of, let's say, cross validation or whatever. Now, I haven't said what's my model is going to be because I have 10 variables and I may pick 3 or 4. Is that okay, or do I have to say, no, I'm going to actually relate the response surface to age and sex and weight, but not to a million other things that I might measure because then we start to get into this world of my making it look like perhaps I can predict perhaps better than I really can.

What procedures do I need to check on how good my predictions are? One of the things that's nice about intention-to-treat and all that is it's pretty tough to game it. So, I think that's a serious issue. When we start to say, let's get more information, it's a serious

issue to say how should we set up the procedures so that everybody is reassured that nobody is cheating anybody?

DR. RUBIN: The point is not the complexity of the procedure, but as long as it can be prespecified in such a way that anyone can duplicate it, that's where its validity comes from. Validity comes from the ability for a robot to carry out the analysis.

DR. LIPICKY: The way in which I understand it -- and Tom can correct me if I'm wrong -- in a sort of morbid/mortal trial, the degree of specificity that is required up front is really pretty serious. It's the specific thing. It's the specific analysis. If you are going to include covariates, you must specify. If you have not done all of those things, then you are looked at with jaundiced eye.

So, what's being talked about now is really quite deviant from that, and maybe if it's descriptive and viewed as needing confirmation, that's okay. I think that's sort of what's being talked about. I think that that's right, and I'm not sure I'm articulating it properly.

DR. PACKER: But I think that clearly to the extent that this kind of information would be deemed desirable and that methods are elucidated in order to obtain this information, then the quality controls by which

these methods are applied need to be clarified so that sponsors are not disadvantaged by trying to do the right thing and finding that a database is modeled by them but remodeled by the agency in a way that reaches a conclusion which changes the whole way that the dose-response curve has been defined.

In other words, I could foresee a circumstance whereby in the learning phases Lew has described, the response surface is described, and then they go and do some specific additional studies in order to either confirm or clarify the response surface, as Lew has described. Then the NDA is submitted and the division then says, well, we looked at the data from the trial which defined your response surface from which you then determined everything else you were going to do in terms of subsequent development, and we just think you got it wrong.

DR. LIPICKY: There isn't any way you can prevent that.

DR. PACKER: Right.

DR. LIPICKY: We looked at black/white and we looked at the U.S. versus the rest of the world.

DR. PACKER: I understand. But my sense is that because of the unfamiliarity that people have with this process, unfamiliarity breeds fear, and one of the fears that could be brought to the table is how do we know

that we're doing the right thing, one.

Over the past many, many years, there has been a narrowing of the number of possible ways one can interpret and analyze a randomized trial done in a conventional manner, and there are certain analyses that are what may be called beyond the acceptable range, which may be done but are not seriously applied.

Right now we don't have that experience here, and one of the fears that may be brought to the table is in the absence of that experience -- and that is an experience that will evolve and get better over time -- what responsibilities are placed on the sponsor and what responsibilities are placed on the division in reviewing the data that allow people to feel comfortable that the conclusions reached will be concordant?

DR. SHEINER: I think that's a very serious question because one of the problems I think we all perceive now is that there's a certain inevitable incentivization for ignorance in the drug development process. We heard about it earlier. We don't want to try with that much higher dose because we might see something that is going to get us into trouble. Well, any system with an incentive for ignorance is a system that's going to get in trouble, and we're going to get that trouble eventually, but we won't know about it. We want to, if at

all possible, make it a positive thing to find out as much as you can about your drug by the time you submit your NDA, within the limits of the resources you have to put to that.

I think what you're getting at, Milton, is that as we allow for much more complex sorts of things, if it's an adversarial relationship, which to some degree it is -- and I was suggesting earlier -- there are more places for somebody to pick a hole in what you did and you've got to know that holes picked in this part are not going to stop me from getting approval. They're just going to be a negotiation process for the label, and is it going to say 15 milligrams or is it going to say 10 milligrams or whatever, or is it going to have this graph in it or that graph in it, but that the approval at least is something that's separate, related to certain confirmatory evidence that you need to have, and that the rules for that and the complexity cannot impact on that. I don't know how to work that out. I know it can be, but I think a concern with it is completely well placed.

DR. PACKER: I think it's important to be up front about recognizing that fear because fears provide a major disincentive here. What we want to do is align the incentives so that people feel motivated towards the same goal, although the motivations may be different, but at least the incentives be aligned.

One of the concerns here is that if there is a major confusional component as to how this can be approached, that that would be a disincentive. Somehow some discussion or a clarification or initial guidance needs to be provided.

DR. LIPICKY: But I think the two can be accomplished at the same time. For the trivial example, one of the designs I had was basically parallel group, three-arm trial. So, there you've got standard stuff. So, that first arm could be said -- you know, and then the other stuff could be treated however one wanted to. Then the only problem would be then the company would ask our statisticians, well, do we have to pay the price of multiple comparisons for our parallel three-arm trial, and I think we would say no. You use that just the way you usually do, and we'll consider the rest of this stuff descriptive and then, depending on how you use it, in later development, find something that would be confirmatory or something along those lines. So, I don't see any problem at all with integrating the two approaches in a practical development program and not penalizing anybody for adopting the approach.

DR. PACKER: Carl?

DR. PECK: I agree there's no problem in doing it in principle, but I think actually providing guidance on

how to do it is really an issue here.

One thing that I'm impressed with in looking at protocols for phase III trials or even -- let's go back -- for phase IIb trials, dose-response, exploratory trials is the lack of information in the statistical analysis section about what will actually be done with the data. Let me clarify.

The phase III trials typically are quite clear about what will be done. It will be a frequentist hypothesis testing thing. So, there's no issue there.

In the IIb trials, the dose response, there's often a statement, the data will be analyzed for dose response. End of sentence. End of section. So, it's left up to the reader to guess what will actually happen. In some cases it's nothing. In other cases it's extensive. Then there's the concern that, well, they looked for what they wanted to look for. They didn't tell us about all the other things they looked for and didn't find. So, there's a problem.

Our center, in collaboration with Lew Sheiner and many others, has been working in the area of modeling, anticipating its use in simulation of clinical trials for the purpose of improving the design in future trials. Realizing this problem of the lack of standards, if you will, for the process of modeling, we have put forth on our

web page a draft document called Simulation of Clinical Trials: Good Practices. It's an attempt to define the principles and the good things to do, if you want your modeling and simulation to be understood and credible.

We've identified three principles. One is clarity, which means that the models and the assumptions and the procedures are all laid out in advance in a simulation plan.

The second is completeness. In other words, you're urged to be complete about that. It's not just a one-sentence thing. We will model and simulate this.

The third is parsimony which has to do with a certain quality of the models to make sure that they're not over-specified.

So, my contribution on this point is that I think if the agency wants to encourage modeling of dose-response data, I would recommend that it engage then in creating some standards or principles of the process so that there's some prospective rigor in what will actually be done. I think that will make it more credible. It will certainly make it reproducible.

DR. PACKER: It will make it reproducible, but I guess one question that arises is, to what degree can one realistically do this?

DR. PECK: I can speak to that. It's not easy

because modelers are like artists. They have their favorite models. They have their favorite procedures. It's an area that needs work to get experts in exploratory modeling to agree on systematic, standard procedures that they're all willing to declare in advance and follow.

But there isn't actually a complete chaotic universe of options here. For example, in modeling dose-response trials, there really are only a handful of plausible basic structural models. There's a linear and there's a nonlinear, and there's usually only one or two nonlinear models that work. Then it's a matter of incorporating the covariates. That's were it can become much more complex and arbitrary. But still there are thousands and thousands of models that are probable. There might be a few tens, and those can be declared in advance by and large.

DR. SHEINER: Can I say one more thing?

DR. PACKER: Lew, yes, please.

DR. SHEINER: First of all, remember right now we don't have an adequate procedure for finding doses. So, anything that gets added on that attempts to do that should not represent a penalty. I think that's got to be a principle.

But the second thing is there's an interesting interaction here, which is that the tendency for the

pharmaceutical company would be to minimize the number of covariates that actually have to be taken into consideration because that makes using their drug more complex. So, you reverse the usual situation. The public, as represented by the FDA, would be inclined to want to watch out for age or other things. Now, as they get to more and more covariates, the certainty will go down. The regulatory agency will be on the side of less certainty because they're worried about safety.

Now, if you can prevent that from being accompanied by a demand for more data, which is what the pharmaceutical companies are going to be afraid of -- every single thing we did in this study, which was more inclusive, as Tom talked about -- now we had four people who were a little bit older, and suddenly maybe there's maybe a little hint of something and it's got low probability, but nonetheless, now the agency says I've got to study 30 old people. That's what you want to avoid.

So, it seems to me that you're going to have to have rules on both sides, but if you do, then it will sort of naturally work itself out because as you get more and more fine, the certainty goes down and nobody will want to make that sort of a claim in the labeling.

But I think the basic principle has got to be, if this is an add-on, at least for some period of time, you

never get hurt if you try to do a better job.

DR. PACKER: That sounds like a good principle.

DR. LIPICKY: Just one more thing to add. I guess I really would like to see, at least in the hypertension area, dose-ranging trials not be called phase II anything. They are phase III. That is the basis of approval. You really have to discard this business of phase I, phase II, phase III. It is where that thing fits in the necessity for getting approved. So, the program in its entirety I think is a different issue, but a dose-ranging trial can be considered phase III, is phase III, and is actually phase III for a combination product. It is the only trial that is required.

DR. PACKER: Lew, let me just ask one other question. The approach that you've described would seem to have lots of the aspects of -- actually it's not the approach you described, Lew. It's the approach that Ray described which, in fact, you might utilize in reaching conclusions about dose response.

One of the issues that we would normally worry about in interpretation of dose response is the issue of carryover effects because when one force titrates but still wants to look at the relationship between dose and effect at any individual dose, in a parallel group design, you don't worry about how you got there, but in an incremental

dose-response design, as Ray outlined, you worry a lot about how you got there. In general, the agency has seen lots of examples where crossover trials have not been easy to interpret because of carryover effects. Yet, the trial design that Ray has put forward has more carryover effects than the typical design. How much should we worry about that?

DR. SHEINER: This is a very complicated issue, which I think maybe Don and Tom could talk more about. First of all, there's again the issue that you have to be within a scientifically valid context. That is to say, the standard model for carryover and period effects and everything essentially says we don't know. Anything could happen. So, you could have any kind of carryover, you could have any kind of period effect, essentially proliferating parameters, and so the problem in figuring out whether they're there or not becomes a big issue and a lot of the tests for whether or not carryover is there are logically inconsistent, in that you fail to reject the null hypothesis and therefore you accept it. And there are all kind of problems like that.

But if you are scientifically based, if you're willing to do that, then you say, well, what's the nature of carryover? Sometimes it's pharmacokinetic and it's pretty simple. The drug persists.

In the case of an escalation trial, the problem is that time and dose are confounded. As time goes by and more time in the trial goes on, that's one thing that's happening, and also dose is rising over time. So, is it the dose or is it the time? It's sort of a way of thinking about carryover but in a more general way.

Of course, there are things you can do about that in terms of the design if you think there's a good reason to believe that it might to take a month or two to develop the full effect of a given dose. One of the things is you can intersperse placebo periods in your escalation sort of randomly. So, you can protect yourself against the things you're worried about, but you've got to make the list have scientific credibility. The things you're worried about have to be as credible scientifically as the things you're trying to claim, and then it becomes tractable.

If, on the other hand, you say I have to protect myself against all possible weird things happening, then there's only one thing you can do, which is a very, very standardized trial, draw a limited set of conclusions, and not even know whether they extrapolate to the future because you've taken such a non-mechanistic view.

So, there's always this tradeoff. The way I think we'll get around it is by recognizing which of the

things we need high certainty for and which of the things that we're willing to accept assumptions and lower certainty because we need to give people a best guess as to what to do and it's never going to be for certain.

DR. PACKER: Jeff and then I want Tom to address that as well.

DR. BORER: I have two sort of general comments and then a question.

DR. PACKER: Don, if you want to sit down, it will be probably easier.

DR. BORER: First of all, I just want to support what Lew said a little while earlier. Much of the discussion has gone on as if the goal is to design the perfect development program and anything short of perfect is a failure, when in fact what we have now is at sort of the lower end of good. If we get to the middle end of good, that's a tremendous advance. So, it doesn't have to be perfect.

The same thing is true with regard to the covariates. We don't ask for information to the nth degree about every possible confound or every possible subpopulation. We don't do that now. It's unrealistic, can't be done. Why are we talking as if it has to be done in some new strategy?

Having said that, just with regard to the

carryover effect, I want to ask a practical question, and Tom or Lew or whoever can answer this I'm sure. In Ray's multiple-arm, forced titration scheme, at the end of forced titration 3 in one arm, you actually were at the same dose as you were in the beginning of another arm and you're 9 weeks later. Could you not look and see that the pattern of effect was the same at the end of forced titration arm 3 as in titration 1 of the other arm and say, we've accounted for carryover? Could you or couldn't you?

DR. SHEINER: It sounds good to me.

DR. PACKER: I guess the answer is possibly yes.

DR. LIPICKY: Yes, but then Tom is going to say, is that a non-inferiority kind of a comparison? Aren't you, Tom?

DR. FLEMING: That's what you'd be doing.

DR. LIPICKY: Yes, and that could get very hairy. At first blush, yes, that would be good.

DR. PACKER: Tom, do you have any additional points?

DR. FLEMING: Actually I've been just trying to gather in all the perspectives here following Donald's talk. Actually I have a few global comments that might be my attempt as a frequentist and someone who does believe in a role of hypothesis testing in confirmatory trials and yet

trying to integrate into that philosophy the insights that came from Donald Rubin's talk.

He made a number of key points. Just a couple of those key points are the issue as to whether a randomized trial itself is giving us a generalizable conclusion to the future because, in fact, if it doesn't, it really isn't serving the purpose that we intend.

He also pointed out the importance of being able to address issues, in my words, that are truly the clinically relevant ones to individuals in trials, and that in fact, involves what is the right dose and schedule for me given my specific background.

I guess to address those issues, the first point that I would make is that it's very important to design trials in ways to achieve maximal generalizability. There are a number of ways to do that: inclusive eligibility criteria; bringing people in from multiple centers instead of having a single center trial; designing the control regimen to represent truly standard of care and to allow a real world access to ancillary care, so that the answers that we're looking at are, in fact, as relevant as possible to the real world setting for the use of your therapy; adequately sized trials to be able to provide, even if it's secondary analyses, relevant insights into exploratory analyses that might be looking at what is the

evidence for homogeneity of effect across various characteristics of patients; providing what evidence is still necessary to be obtained on dose and schedule, although I would argue, as I did toward the beginning of today, that that's an issue that we should be addressing from the very beginning of our drug development plan. And we do. Preclinical studies, PK studies back in phase I, in particular, looking at an array of doses in phase II and IIb, and in fact, in some cases as well, looking at these issues in phase III.

So, I would hope that even though one can never fully achieve it, that one is designing trials that are confirmatory trials to be as generalizable as possible and as relevant as possible to the real world setting.

When one gets to the confirmatory trial, i.e., when one gets to what I call the phase III study, it is in fact important from my perspective that we have prespecified goals that we are addressing in a confirmatory fashion, and for multiple reasons, including those that Lew articulated, that there are a myriad of different exploratory analyses that can be done, many of which certainly lead to what some of us might refer to as data-driven aspects and it's difficult to discern level of convincingness.

Essentially ultimately what I want to be able

to assess is the strength of evidence. I wish to have a standard for that strength of evidence. To my way of thinking, the paradigm that we follow at this point for prespecifying key hypotheses and providing hypothesis testing gives us one effective way of being able to address whether or not there is adequate evidence here to meet our standards for strength of evidence to establish benefit and to establish a safety profile that gives us a favorable benefit to risk.

That doesn't mean, however, that these methods can't be adaptive. There certainly are, depending on our level of confidence of how much we're willing to prespecify, valid methods that allow flexibility in the formulation of these statistical methods that are used that still allow us to meet this frequentist paradigm and p values and standards for strength of evidence.

But certainly we don't stop at that point. If we did, you would only need a statistician to interpret results and make judgments. As important as some of us might believe statisticians are to this process, obviously this is only one part of the process and a global judgment, looking at all of the results in the trial, strength of evidence not only of what we see in this primary hypothesis, but secondary measures, safety considerations and results from external trials, is critical and needs to

be integrated in a way that is obviously subjective and requires considerable judgment.

In this process of designing trials, to come back to the specifics here of dose response, I would strongly support the argument that phase III can still be a setting -- and maybe should be a setting in some instances -- for looking at dose response and looking at multiple doses.

Being a frequentist, I worry about the issue of multiple testing. Somewhat unsuccessfully, I've been arguing for a long time that multiple testing adjustments, though, shouldn't apply to this specific issue of looking at multiple doses against a control. I've had multiple long discussions with FDA colleagues and others about this issue, and I know I'm in a minority in the FDA ranks about this. I believe in adjustment for multiple testing. If you are a frequentist and you're looking at strength of evidence, if you allow multiple endpoints, multiple test statistics to look at those multiple endpoints, testing over multiple points in time, looking at multiple subsets, all of these greatly inflate the risk of getting differences that would appear to be significant evidence of benefit when, in fact, they could be attributed to random variability.

I view, though, in a very different way looking

at multiple doses, and the irony is if someone was so naive as to say, I'm going to look at low dose against control in one study and high dose against control in another, and not be asked to adjust for multiple testing, that's logically inconsistent with that they had the wisdom to say I could be more efficient to put high dose and low dose against a common control, that I'm going force them to adjust for multiple testing.

Or to say if low dose against control gives a two-sided p value of .04 and high dose control is a two-sided p value of .04, that's no longer significant because it doesn't meet some kind of multiple testing adjustment. I've actually strengthened the argument when I see high dose against control being .04 when I also see low dose against control being .04.

So, I've argued that if one does look at multiple doses against a control, it is very appropriate to control the pair-wise comparison error rate, adjusting for all multiple comparisons done within that pair-wise comparison, but one should factor in the strength of evidence from the multiple doses using some kind of aggregation, whether it's a dose response or some other method, some type of aggregation for strength of evidence.

Now, one last comment and that is should we be doing a primary analysis for efficacy. Let's say we fit a

log-linear model or a linear model and the slope is our strength of evidence for our primary hypothesis. There certainly can be some ambiguities that can result. In fact, one concern that I have is if you look at high dose and low dose and there's no control but the low dose is, in essence, thought to be the control for the high dose, if there isn't a dose response, I'm left without certainty as to whether the two doses are equally effective or equally ineffective.

If I have a control in a low dose and a high dose and I'm fitting a slope and I find a significant relationship, but I then also discover that the high dose has an unacceptable safety profile, I'm uncertain from that analysis whether the low dose, which has an acceptable safety profile, actually has had an established benefit.

So, there are some complexities with these types of analyses, and I have tended as a result for this reason, even though Carl might call me inefficient, to favor designing these phase III trials in ways that do allow me to obtain adequate sensitivity to each dose against a control so that if there is, in fact, significant evidence or true, significant benefit for a given dose, I'm able to address that and have sensitivity to it.

But I would definitely take this opportunity to argue one more time that if someone has the wisdom to look

at multiple doses against a control, that they ought not be paying an alpha spending penalty for that type of multiple testing.

DR. PACKER: I think for the sake of time and the sake of clarity, I think we heard Ray say before about an hour ago that you would not penalize a sponsor who wanted to undertake this path and exact a penalty for multiple comparisons, that you would allow each of these arms to be tested pair-wise at an alpha of .05. Is that correct?

DR. LIPICKY: See, I don't understand the thinking. I agree 100 percent with what Tom said, every single word. I think I agree with what you said, Milton, but I don't see how anything that either of you said is applicable to this circumstance. This circumstance of an antihypertensive drug in the current development regimens have no phase III. There is no such as a phase III. There is no such thing as an intent-to-treat, morbid/mortal trial where all these rules start to really play enormously important things.

So, I think what I was saying is I wanted to see a well-defined shape of a curve somewhere -- and maybe that could be done first in a learning phase and then a second trial in a complementary -- whatever the word is -- "you prove that your first learning was right" phase. Then

that defines the dose response. I think you could do that in 300 patients. Now you've got another 2,000 patients you can devote to other things to learn about the antihypertensive. All I want to do is consider at this point in time what you can do from the dose-response point of view, if we could settle that.

At the moment we have 6 and 8 trials of 300 to 500 patients each that are all parallel, fixed-dose, dose-ranging trials. That's the whole NDA. In fact, every one of those trials is treated by a frequentist type analytical process where something is being tested against placebo in a pair-wise comparison way or sometimes ANOVA, sometimes with some kind of more fancy thing. But I think that's just all the wrong stuff. So, although everyone is saying good stuff, it isn't applicable to the circumstance that basically needs to be changed, which I think it should be.

DR. PACKER: I think that speaks to the point that you made before that you made before that if one knew that the response surface would look like, one can reach conclusions independent of the calculation of a p value.

Tom, my sense is that it's not either/or, that these are complementary approaches that enhance the quality of the database.

DR. LIPICKY: I don't disagree with Tom. I think he's 100 percent right. It's not addressing the

right problem.

DR. PACKER: It sounds like this is much easier than the Middle East.

(Laughter.)

DR. KONSTAM: With your construct of what you're talking about in terms of not having to pay a multiplicity penalty, I assume that that really implies a very strong assumption regarding something about the model of the dose response. Let me give you the ultimate which works with hypertension because I think we know a lot about dose response in hypertension. The ultimate contrary example would be what would you do with a drug that you knew nothing about dose response, including the possibility that higher doses may go in the opposite direction. And you may do placebo and two doses and find a low dose looks like it reduces mortality and higher doses actually looks like it kills people, which has actually been seen. So, I guess it assumes that it fits some model of dose response, doesn't it?

DR. FLEMING: Well, what I'm assuming, when I'm arguing that I don't do alpha adjustment, is if I have set up the study to address multiple schedules, which could be multiple doses. My primary analysis when I'm looking at it as a frequentist here for whether I hit the standard for strength of evidence, establishing efficacy, is to look at

this in a pair-wise comparison of a single dose against control.

When, however, I look at the totality of the data, I will do some type of meta-analysis. That meta-analysis may be multiple studies looking at that same dose, or it might be combinations of studies that look at multiple doses against a control. Clearly then there's going to be in that level of meta-analysis, where I'm aggregating strength of data, some type of model. When I'm aggregating data from multiple sources of information, I will then be using some type of model.

For example, if I had two trials looking at the same dose and one study showed a favorable effect and one study showed an unfavorable effect, the unfavorable result weakens my confidence in the result in the favorable. If, in fact, I'm looking at multiple doses, yes, I'm certainly going to either formally or informally, intuitively, be fitting some kind of model that I'm assuming for dose response. I'm happy with either one depending upon which approach you would prefer. I will then use that model, formally or informally, to integrate all of the insights that I have to determine, A, whether there's evidence of benefit and, B, is there evidence to suggest what is the dose that maximizes benefit to risk.

DR. LIPICKY: Really, you need a lot of

particulars. I think both of the conceptual aspects of what we're talking about can be satisfied by a single design. It would take a lot of thought and careful thought, and I don't think one has to be dichotomous and say we can't do one or the other. I think both can be done at the same time.

DR. FLEMING: Ray, admittedly I'm having trouble focusing my thoughts on this to the single case of antihypertensive drugs where we're going to assume that we can rely on a surrogate in this setting, i.e., blood pressure lowering, and we're going to assume that there is no dose-limiting toxicity. I'm really thinking in a more global sense.

But the principles that I've laid out, if applied to this specific setting, would argue from my perspective that if I did believe that I had no dose-limiting toxicity, that it would be very prudent not just because of what the FDA would wish, but from the sponsor's perspective to set up a wide range of doses because this is an answer that I can get for a wide range of doses.

If you are looking at survival as the endpoint or stroke or MIs and it takes 10,000 patients to be able to address the effect of a single dose, those studies don't lend themselves as well to understanding the dose-response surface.

But if I can rely on a surrogate and I can understand the relationship of the effect of a given regimen on that surrogate endpoint in much smaller numbers, and now you're telling me I don't even have to worry about whether there's an upper limit of how far I can go because of a dose-limiting toxicity, then I would argue it makes perfect sense to look at, as you've advocated, a wide range of doses that allow me, with a reasonable sample size, to get an insight into the dose-response surface and to fit some model, if I'm a frequentist and want to look at strength of evidence, into how I'm modeling that relationship to establish whether I have conclusive evidence that the intervention is affecting this specific outcome.

DR. SHEINER: I think that this issue of a single study being used for two purposes -- And, Tom, I admire your pure position of integrating it all and having a an overarching view of it. But let me segregate it for a moment. Let me propose something that's been proposed in the past.

There are some people who really want to see something clearly demonstrated by the data, so I'm going to do two doses: placebo, which is a dose, and a dose that I think will work. I'm going to do that in 400 people: 200 placebo, 200 with this dose. And I'm going to do it for 1

month or whatever it is. That's what I'm going to use to demonstrate that the drug works.

Now, in that exact same study, I will, at the end of the 1 month, randomize all of the people to some other dose. So, I'll now actually have a second period in which they'll all get some other dose. And now I might choose for just that one dose that I used in the first period.

Secondly, from the very front of this study, I'm going to add on another 200 people who are much more inclusive. They're not having blood pressure exactly where I wanted them to be in the study. They're not without any other complicating disease. I might take people with angina when I didn't before. And I put them also in this same study.

I can do two completely separate analyses of the core of 400 people in a straight-ahead, hypothesis testing way. Then I can take that data, add it on to all the other data, and be a complete Bayesian, for example, and just do a total model-based approach with informative priors designed to come up with the best understanding I have of the response surface. There's nothing illogical. There's nothing wrong. The fact that I got all that extra data in that same study does not mean that that analysis of the core 400 all by themselves, having stated beforehand

that that's the way I was going to do it -- it doesn't change in any way the frequentist properties associated with that and the conclusions we might draw from it. So, although obviously it's in some sense more satisfying to look at the whole thing as all one thing, you can do that, what I just said, and that would be an example of getting extra power for the exploratory analysis and no loss of power for the confirmatory one.

DR. FLEMING: In fact, I would argue that's precisely the approach that we should be taking and, in general, are taking although aren't necessarily doing it with quite the level of rigor as you've explained for the exploratory analysis.

The approach in general that I would advocate is to design trials to be generalizable, to be relevant to the real world setting, and to be able to be analyzed for strength of evidence from a frequentist perspective, but then ultimately the assessment of benefit to risk and approvability and as a consumer, whether I'm going to use a given agent, should be based on an aggregation of all of the data that will clearly have to involve an exploratory element of this.

A Bayesian approach is one relevant approach. If I want to use a prior and get a Bayesian inference, that's a very relevant exploratory approach. Of course, we

have to understand the influence of how my choice of the prior would influence the interpretation, but that can all be laid out.

DR. SHEINER: Let me just add one thing to that, Tom. Having said all that and we're in 100 percent agreement, my extra little claim that I'm not sure whether you're for or not or what you want to say about this is that that explanatory -- I like to call it explanatory rather than exploratory because it makes it sound a little better -- analysis, having satisfied the frequentist test, might show up -- the conclusions of that might show up in the label as a graph of dose response or as a set of suggestions for doses or something rather than the label simply saying we tested, in a standard confirmatory way, 100 milligrams, so that's what's suggested.

Rather it might say now, given that 100 milligrams definitely works, this is where it appears to be on the dose-response curve and just interpolation, this is what the picture looks like. So, if your patient has these characteristics, you might consider lowering the dose to that and so on and so forth.

DR. FLEMING: I guess I would be unwilling to make an absolute statement as to what would always be the case. It strikes me that the possibility that such insights could be very important beyond what came from the

primary prespecified frequentist analyses certainly exists, that those results could be important additional insights and the reliability of those analyses sufficiently established that it's possible.

DR. LIPICKY: So, we just have to convince somebody that it is worth their undertaking these things so we can get some data set to examine and see whether anything we're talking about is real or not.

DR. PACKER: Well, I think that brings us to the final part of today's agenda. I'm going to take the chairman's prerogative to skip all of the questions.

DR. LIPICKY: Okay.

(Laughter.)

DR. PACKER: I want the committee to address two questions, and we can do so in detail or in a general manner, depending on what the responses are. But I think the two questions that summarize -- I mean, what we have heard is that a problem has been identified in terms of the adequacy or inadequacy of characterizing the relationship between dose and response in the evaluation of drugs for the treatment of hypertension. An approach or, should I say, approaches have been proposed that would greatly expand the knowledge base, and this approach does not conflict with our conventional approach to hypothesis testing. The two approaches are complementary, and that

this additional approach does not constitute an undue burden on sponsors and, in fact, may alleviate them from the responsibility, expense, and time of doing conventional analyses that may contribute little to the total knowledge of what the appropriate relationship between dose and response may be. Therefore, the new approach would provide benefits to not only industry but to the regulatory process and, most importantly, to patients. And we've outlined many of those reasons.

So, that leaves us really with only two questions. One question is, are there reasons that sponsors should, in fact, elucidate the full range of the dose-response relationship or, as Lew says, the response surface? Are there reasons?

I think we've elucidated many, many reasons why it's in the sponsor's interest, in the interest of the information gained, the clarity of labeling, the clarity of the development program, and in fact in the public health interest. We mentioned some of those issues earlier. Especially as the target blood pressures decrease, the number of drugs would increase unless the drugs that we have had a wider dose-response relationship than previously assumed.

So, the last question to this group is, are there reasons that sponsors should not elucidate the full

relationship between dose and response? Are there reasons that they shouldn't do that? I guess I'd like to hear any proposals why they shouldn't.

DR. KONSTAM: I still think there are reasons why they wouldn't.

DR. LIPICKY: No. Why they shouldn't.

DR. PACKER: We have outlined several, maybe six or seven, reasons why they should. Are there reasons why they shouldn't?

DR. BAKRIS: Milton, let me just make a comment to that. Marvin is right. There are reasons they wouldn't, but why they shouldn't, the only reason I could imagine would be an economic one, and I think we've made a pretty compelling argument that, in fact, from an economics standpoint, it's in their favor. It probably is more cost effective. So, you've actually stumped me with that question. I usually can come up with something, but I can't think of anything for this one.

DR. PACKER: The question is framed specifically with intent. The word "should" is in fact the intended word. Are there reasons why they shouldn't? If there are many reasons why they should and there are no reasons why they shouldn't, then they should.

DR. BAKRIS: That was subtle.

(Laughter.)

DR. SHEINER: The same compelling logic as hypothesis testing.

(Laughter.)

DR. PACKER: Consequently, we've outlined reasons why they would. So, if they should and they would, then they might.

(Laughter.)

DR. PECK: Well, I'm going to rise to the occasion and try to anticipate a shouldn't. That would be under the circumstance that the prospect of harm to a subject, even remote, would be unacceptable to a company. That is, their view that having one adverse reaction of a particular kind, say, a teratogenic reaction, would be so devastating to their image or reputation and the drug's image that they would be unwilling to creep too far into the unknown.

DR. LIPICKY: I think there's a very good answer to that. Much better that that should be in the development phase than postmarketing. So, that just is not an adequate reason. People aren't going to get sued in the development program. They're going to get sued postmarketing.

DR. PECK: Yes, that's all right, but small companies whose finances are subject to private financing or mouth-to-mouth, month-to-month, or flaky stock market