Every one of the folks presenting from the patient representatives talked about how important quality of life was from, again, this gestalt aspect. And I think this is where my kids come in in terms of quality of life is kind of what I'm doing this for. Also, I must admit to a certain amount of guilt. They're at Disney this morning, but I had to leave them for this particular meeting.

(Laughter.)

DR. SLOAN: So, you can see a little bit of frowning on their faces. That's what it's for.

Thank you.

DR. CELLA: Thank you and thank you from tearing away from Disney World. You can at least tell them that you just showed them to the country.

Dr. Nerenstone.

DR. NERENSTONE: I just have some very brief comments. Actually I approach this a little bit broader and was both glad and horrified to find out that my outline very much paralleled Dr. Schilsky's outline. So, I'm really not going to dwell that much on a lot of the same overview points.

But listening to Dr. Sloan, I was very much taken by how much he and I agree on a very fundamental idea, that whatever instrument that we are going to be using or that's going to be developed or that's proposed to

us needs to be interpretable by the clinician, but it needs to be simple. I really, really agree with that. Maybe it is because we clinicians are just simple as well, but I think that for the results to be believable, to be interpretable, and to be useful clinically, they need to be very simplified but useful. And I agree with you, I think they can be made that way.

I think we have to make sure that we can try and get some standards for all studies. And to this, I ask the FDA if they would consider involving the clinical cooperative groups as well. We know that really they're important in clinical trials and drug development as well, and if we could get some uniformity of definitions, it would make it much easier for the clinicians who are putting patients on study to actually get the information. I think this is going to come out about lack of information. Especially when you have something that's going to be looking at big shifts as being important, then missing data is going to be even more destructive to the integrity of the study. So, if everybody is sort of on the same page, it's going to be much easier for us to compare trials, to compare studies, and to have the doctors do what you want them to do.

One of the things that was mentioned briefly is quality of life as an endpoint in itself versus quality of

life as a secondary endpoint. I think that will probably very much be specified by the drug and by the phase of development that that drug is at at the particular time.

But I would urge, even when we're not talking about quality of life as a primary endpoint, that the FDA recommend some of these parameters being followed even in their earlier development. That's because this is going to be used as a tool by clinicians. Clinicians are going to have a choice, even assuming that a drug is available for use because of response activity in a certain disease. It's going to be very important for the clinicians and their patients to make a decision whether to use the drug based on some of this data that would be important to have, not only toxicity profiling, but some of the more subjective things that we're talking about.

Separate from the discussion that's gone before, I just have one other point that really hasn't been brought up yet, and that's the effect of what I call the placebo effect or the investigator bias that no one has really brought up as being a problem in this whole area of quality of life. I think most people would agree that even symptoms can very much be influenced by the act of taking a pill or being involved in a trial or taking a drug. Investigator bias is extraordinarily difficult to quantify. Even in your randomized trials, drug A is the standard, but

you've been randomized to drug B, and this new drug really looks very promising. How is your pain, Mrs. Jones?

I think that this very subtle influence of investigators, which all of us have because if you're involved in clinical trials, you want new drugs to succeed, you want things to be better, is really not talked about. And I think very important when you're going to be having drugs that potentially could be licensed because of their effectiveness on symptom control, in fact you may be seeing a very strong placebo effect.

That's it.

DR. CELLA: Thank you.

Any comments, discussion?

DR. SCHILSKY: Dave, I think one of the most important comments that you made, Jeff, has to do with the interaction between the clinical investigator and the quality of life investigator. I think it's critically important that, as trials are designed, that the two investigator groups sit down, get together, think carefully about what are the clinically important parameters that should be measured, and then figure out a strategy to how best to measure them.

Of course, I completely agree with your statement about the insignificance of a significant p value if it's not rooted in some hypothesis about a clinical

effect. This gets back to one of my comments earlier about the importance of trying to have hypothesis-directed research in this area. We can't always anticipate what the effect of a treatment is going to be, but oftentimes we can at least make a stab at it and develop a hypothesis. I think that that really has to underlie a lot of the way we design trials with respect to these sort of nonmedical outcomes, if you will. So, the interaction at the design stage between the quality of life investigator and the clinician I think is critically important.

DR. CELLA: By the way, I remind you to turn to the second page of the Points to Consider handout. The audience has it all on one page. The subcommittee has more work and so we have it on three pages. So, for the subcommittee, the second page.

This really just goes back to Jeff's presentation and can focus some of our thoughts. What are optimal and minimally acceptable responsiveness data -- for example, effect size, significance testing, which we've already heard about, et cetera -- that could be used to assess group comparisons? And then what are the acceptable clinical or statistical approaches for assessing the magnitude of change in individual measurements?

We've talked about these things, but I'd like to focus in on this group versus individual comparison

issue and then, with the remaining time, later to discuss the amount of supporting evidence that would be sufficient to allow clinical interpretability of questions and summary scores.

So, about this responsiveness in groups comparisons versus individual, Dr. Sloan had a few things to say there. Are there any subcommittee comments or perspectives you'd like to start with?

DR. MOINPOUR: I'd just like a point of information. In FDA deliberations about applications, do you commonly look at the issue of individual differences, or are you looking at group comparisons primarily?

DR. CHIAO: I can try to answer that. For the symptom palliative endpoints in the two prostate cancer drug trials that we presented to ODAC, mitoxantrone, prednisone, and suramin, I think the first one has the individual changes in terms of individual patients and characterized as responders versus nonresponders. And the second trial, suramin, the prespecified analysis actually is the comparison between the mean pain score across the two groups. But we did the exploratory analyses looking at the responders versus nonresponders. So, the answer is yes, we've looked at individual patients.

DR. CELLA: Stacy?

DR. NERENSTONE: I think we're even going to

have to take a step back a little bit and say is the FDA willing to say to drug companies, drug developers that they're going to need to increase their sample size to start accruing patients who have symptoms to begin with. That's sort of going to be sticky because we know that if you're going to impact on survival or even response rates, you usually want the best patient population you can because if the patient population is too sick, they're not going to respond. We know that. Multiply treated patients, patients who already are performance status 3 and 4 are much less likely to respond.

So, then you're going to get some flack, and I think legitimately, from the drug companies saying that those patients are going to dilute our results. And our primary endpoint is not quality of life, it's response rate or survival.

So, then they're going to say, well, how about if we separate them out, but then there's a question of subset analysis. So, Dr. Pazdur.

DR. PAZDUR: Let me address this issue because it's a very complicated issue and it brings forward comments that Rich and I kind of echoed, that many of the trials that are being done are done in performance status 0 and 1 populations where quality of life determinations in an asymptomatic population may be there, but obviously

analysis of the symptoms and other components of quality of life would be very difficult to interpret in an asymptomatic population.

You could even make the perspective of are the drug companies kind of front loading their studies to have very good patients in them to make the drug look better than it actually would be in a general population because many times when the drug is used, we do not label a drug for only for use in performance status 0 and 1 populations. So, in a sense by having this up-front population of good performance status, we may not be giving an adequate picture of how the drug is eventually going to be used in the general population. And we know from specific examples that performance status can have a marked effect not only on efficacy, but definitely on toxicity.

Given that, one approach that we have taken is the following. Since most of our regulations or, I should say, our regulations and consideration is to have two trials done for an indication, one of the suggestions that we have offered to drug companies is to focus a study on a specific kind of conventional endpoint demonstrating improvement in, for example, survival, then a second study specifically looking at symptomatic patients, demonstrating that this endpoint should be the primary endpoint and that we would want to have this as the primary endpoint with the

statistics geared toward a quality of life determination. This is what we're starting to at least evolve in some of the discussions with the companies.

Because I think, as was pointed out by Rich and the other people, one of the problems that we have is that many of these quality of life tools are added on without discussion with the investigators. They're kind of lopped on at the end. Well, we need our requisite quality of life tool here to make this a kosher study and not a lot of consideration given. So, by really making that, in a second study, a primary focus, some symptom benefit or some quality of life benefit -- and I'm using these words relatively loosely here vis-a-vis our previous conversation -- we can focus attention on this clinical benefit which may not be just a survival benefit.

I hope I answered your question.

DR. NERENSTONE: But what happens if they're discordant?

DR. PAZDUR: Then, as with anything, it's a review issue and we have to take a look. This is true in many studies that we deal with, even when we're looking at survival. We have to take a look at the relative risk/benefit of the drug in these populations, et cetera. So, this is not unique just to this analogy but would be seen even when we're taking a look at survival or more

classical, conventional approaches to drug approval.

DR. CELLA: It's well known that there's a pretty large efficacy/effectiveness gap, if you define efficacy in the usual way of results from a phase III trial that you typically see and effectiveness being what happens when the drug goes out on the market and is used. It sounds like there's an interest in narrowing that gap by directing these two-part studies, if you will, or two-part submissions that actually span the eligibility criteria more broadly. I imagine that the intention is to move that into labeling as well, or is that not --

DR. PAZDUR: This is a point under discussion. I can't make a generalized comment --

DR. CELLA: It's also not the purpose of our committee.

DR. PAZDUR: But it makes sense I think to focus on the population that is going to be getting the drug, rather than making an imaginary best scenario population to be using it.

DR. CELLA: I think it will be useful for this subcommittee to sort of track those discussions as they become public within the agency.

For our purposes, we've mostly I think in this context been referring to individual change and how much change does an individual need to have, the implication

being improving in symptoms, but the other side is a worsening. So, for example, if we can move this field, if you will, toward identification of what's a meaningful change in an individual person on a given metric, given questionnaire, then getting worse on that scale may or may not have the same meaning as getting better on that scale. But just as you might talk about symptom improvement, you might also talk about delay of symptom onset, using the same distance change that needs to happen to define what it means to call it a symptom onset.

DR. PAZDUR: And that I think is particularly interesting when we take a look at the cytostatic drugs as one of the presenters, the patient advocate, presented. When you're looking at drugs that do not classically reduce tumor size, the delay in onset of symptoms may be a very relevant clinical endpoint.

DR. CELLA: So, let's focus, if we can, then on the individual side now. What do we know and what can we say about the best available methods for determining what the meaningful improvement or meaningful worsening on any of these health status, quality of life, functional status, symptom scales that exist and come before ODAC? Lillian and then Diane.

DR. NAIL: My response from our research group would be right now not much. There is a difference in the

size of change that patients feel is clinically significant, depending upon whether they're improving or getting worse. Getting worse is much more noticeable to you. Improving is a little more difficult to figure out, and improving seems to be something that comes to people later.

Many of the instruments that are in use today have not been tested in a situation where we can tell if they're really responsive to a known change, to a place where clinicians say patients on this treatment change in their level of symptom X or their level of quality of life from point A to point B, and we know what the size of that change is. In fact, there is some data to suggest that some of those instruments are actually measures of a trait rather than a changeable state, and I think that's a basic issue that needs to be addressed.

One piece of that issue is the timing of measurement. When we are only collecting data -- and I'm not talking about the cytostatic drugs now; I'm talking about the cytotoxic drugs -- at the time people come back for their next treatment, that's their best point. There is a huge demand characteristic here because many people believe that if they're having a lot of side effects and they tell the person who's prescribing the drug, that's going to get them off of treatment. That situational

interpretation has not been studied very well, but we hear about it in the clinical setting and in the studies.

When we've measured people over time, where we're doing telephone calls or daily diaries, we get a very different suggestion of the pattern of side effects than came out of the studies where the only measure was at the time of the visit. And I think that's a methodologic issue that needs to be addressed. It has huge cost considerations.

DR. CELLA: Diane?

DR. FAIRCLOUGH: The thing that I wanted to clarify was to make sure that we understand whether we're defining a change in an individual that is significant in terms of classifying that person possibly as a responder or nonresponder and then putting that into a group analysis versus the issue of saying that change is significant enough for that patient that we should do an intervention because there's a cost in sensitivity specificity. You could actually in the first case have a little bit -- there's an acceptable error in classifying them in terms of responder or nonresponder in the context of a large analysis. There is a much finer error that we would allow in terms of following up with an intervention.

I think for the purposes of ODAC, the former is defined, but I don't want somebody to walk out of here and

say, okay, we said a 5-point change is significant in an individual and then they could say, we should use that for clinical intervention.

DR. CELLA: So, that's a good point. The issue really is that there's a degree of error involved in this determination, and for the purposes of deciding, within reason, that a meaningful change has occurred in a person, you're optimistic about being able to do that with several different scales. However, it wouldn't necessarily be at the level of deciding an intervention needs to happen for that person.

Julie.

DR. BEITZ: I just wanted to point out that besides symptom improvement or worsening, there's also stabilization, and that many times we're shown data to suggest to us that the patient is no worse than how they started out. I was wondering how you all thought that the effectiveness of the tools that we're using are to showing stabilization.

DR. CELLA: Stacy?

DR. NERENSTONE: I think investigator bias is even worse with stabilization for all the reasons that I said before.

DR. SLOAN: I just wanted to follow up on Diane's comments. I agree totally with what Diane said.

Getting back to what Dr. Schilsky said about the interaction between the clinician and the quality of life investigator, if you will, although sometimes they're one and the same, in assessing a priori what effect we're going to define as clinically significant for a clinical trial -- let's say, if we were to take one of the methods, the errors approach -- it's the one I'm most familiar with I guess -- and talk about a moderate effect size being half a standard deviation on, let's say, a particular instrument, that might mean on a 13-item instrument, each one scaled from 0 to 5, that 6 of the 13 questions will have changed by one category for the entire group. These sort of discussions with the clinicians a priori I found incredibly useful to clarify the issues, as Diane is talking about. What is important to the individual? What is important to the group?

I think again if that interaction between the entire research team is good, you can come to a consensus and again provide appropriate documentation and say, all right, a moderate effect size here is sufficient because we think that changing people an average of one category on 6 out of 13 questions is a clinically important group change. If a person in my office changed on 6 out of 13 items one category, I might not clinically intervene. I particularly wouldn't since I'm not a clinician. As Diane said, the

issue might be, no, I'd want to see the person change 10 points because that would mean to me 10 out of the 13 items had at least changed one category.

DR. CELLA: Do you want to follow up on that, Rich? I'd just like to follow up on that and ask a question.

So, let's say you have this 13-item scale and going into the trial you say that you, the clinician, having talked with your quality of life measurement person, whoever developed the scale or perhaps somebody that's at your local institution, whatever, you two have agreed that if 6 of those questions change at least by one category, which would be a 6-point change in raw score terms, that that's meaningful. Then let's say that 6 points is half a standard deviation. So, you start converging in some evidence and say, now we've got two people that agree that a 6-point change representing at least six areas is a meaningful change.

Is that enough, if that's corroborated by evidence that that amount of change on that scale is about a half a standard deviation and in effect size terms would satisfy something you laid out earlier? Is that enough? Or should there be some other kind of pretrial activity, engaging other clinicians, for example, engaging other patients in this discussion, a look at these questions by

an expert panel pulled together in some way? What do you think?

DR. SLOAN: Yes. That's an excellent question. Certainly two people deciding in a room, yes, this is good enough should not be sufficient evidence for any application to go forward. I think it's a point to start at.

From that, though, given the techniques that we do have, as I mentioned, because the four techniques, whichever way you'd like to justify or examine the question as to whether this change is clinically significant from a statistical standpoint, from a historical standpoint, from the literature of the tool, from the norms that have been published on the tool, I think a sound, scientific, objective justification that, yes, what we think is important is actually within the realm of importance from what the literature and others have told us is a reasonable thing to do.

And then I think it becomes almost case-specific as to how mature, for example, the tool is. If you'll forgive me, let's say, using the FACT instrumentation, I would feel fairly comfortable in going forward with a little bit less evidence than I would with something that had been out for just a couple years and tested on just a few people, for example, or in a very

specific situation because there are normative data out there. There are reliability and validity and good estimates of variability for the tool scores that we can appeal to to say, okay, if what we think the standard deviation is is drawn from the literature and it's justified by what has been published on the FACT instrumentation and this seems to be, relative to what other people have done, a clinically important or observable effect, then again, as long as it's a justifiable, defensible argument, then I think that's reasonable.

The other point I'd want to add to that is, yes, I think in some situations, for example, where a tool has not been used in a particular population, I think a pilot study or an expert panel is definitely a good idea. Again, how you might wish to justify that, make that scientific argument can change with each application, but certainly each element that you mentioned should be there I think in some degree.

DR. CELLA: Rich and Lillian?

DR. SCHILSKY: I guess I just sort of had a question for the group because I don't have much experience in forms development or forms validation. But we concluded the morning session with your first consensus statement that the patient is the expert, and if the patient is the

expert, then it would seem to me that we would want to have the experts involved in designing the tools. So, to what extent are the patients or have patients been involved or are patients involved both in designing tools and in reaching these conclusions at the beginning of the trial as to what amount of change in any given scale is important?

DR. CELLA: It's variable. Some questionnaires were created by so-called experts who represent the patients through their experience. Others are developed almost exclusively by asking patients, and then most are developed with a mix of input. So, for most of the things that you'll see -- and you can always go back to the source publications or request that information -- there was input from patients.

However, it continues to be a challenge for two reasons really. One is that sometimes even though patients are the experts on how they're doing, they don't always have the best view on how to explain the problem in a way that helps you create questions. So, we tend to need input from both patients and providers who are more comfortable with kind of classifying the problem set, if you will.

It's also complicated because the target moves. Disease symptoms tend to remain fairly constant, but treatment side effects change as treatments change. So, there's a constant need to develop new questions to get the

kind of sensitivity to the down side, if you will, of the treatments that are emerging.

I think we need to move around to some people. Lillian, did you still have your hand up?

DR. NAIL: I was going to respond to the issue about patient involvement and the determination of minimally important clinical differences. Our experience has been that the clinicians really don't understand what the patient's day-to-day life is like, and having a clinician make a decision about or a researcher who has no interaction with the patients about what the size of the difference is really doesn't work very well from our perspective.

Now, Jeff had mentioned Jaeschke's work on minimally important clinical differences where they ask patients did you notice a difference and then they look at what the change is the scores would be. The weak psychometric piece of that is that did you notice a difference question. We've used it. We still have some concerns about it, but we think it's better than some of the other things.

And I was just going over some of our data and all of our effect sizes are greater than half a standard deviation. But this is in a symptom measure not in a function measure, and it's a single symptom that we were

looking at. We think that has some promise, but it also has some problems.

DR. CELLA: Carol?

DR. MOINPOUR: Well, I was just going to propose that as a committee that we restrict maybe our eventual recommendations about clinically significant differences to the group level because I really believe that treatment decisions from clinical trials, by and large, are dealt with in terms of group findings, because if you look at the individual variation in patients' ability to metabolize drugs, there are all sorts of things that affect whether or not a particular treatment that's been shown in a clinical trial will actually work with an individual. I don't think we should be any more forced to deal with this for the quality of life data than in the treatment setting.

I think we're attempting to understand that better at the individual level. We know we have to have more reliable questionnaires than we do for group level comparisons. But I would say that, for the time being, we might deal more with group level, clinically important differences at the group level, and not focus on the individual measurements.

DR. CELLA: Yes.

DR. JUSTICE: I'd just like to get back to a

point that was raised by Dr. Nerenstone, and that is blinding of trials is a problem. I think we really need to consider it when we're thinking about effect size. Oncology trials are traditionally very difficult to blind for various reasons. Oral agents are easier to blind, but the parenterals are not. I think the effect size that's needed would depend on the trial design, whether it's blinded or not blinded.

We've taken the position, for example, in the mitoxantrone that a large effect size in an unblinded trial might be believable, whereas a smaller effect size in an unblinded trial might not be. So, that's an additional complication when you're thinking about effect size and what would be needed.

Just another comment is one way to get around that is to try to support an effect in an individual patient by some other objective measurement such as tumor response, and that would be an argument for looking at individual patient responses.

DR. CELLA: Does anyone know if we know anything about whether a placebo effect might be related to either investigator bias, as Stacy described, or other factors that contribute to patient desire to have benefit from a treatment, whether that's more pronounced with symptom improvement or more pronounced with symptom onset?

That is, are you equally concerned about the problem in a trial that would look at delay to symptom onset as with a trial that looks at symptom improvement, and is there any data to support that concern one way or the other?

Dr. Temple joined us.

DR. TEMPLE: There have been a lot of publications recently about "the placebo effect," most of which misinterpret the phenomenon entirely and just attribute the change in the placebo-treated group to be placebo effect. But there have been very few attempts outside of certain specific situations like acute pain to quantify and evaluate that. So, I don't think there's a good answer to your question. I'm virtually sure there isn't.

DR. PAZDUR: I would just look at it as a bias is a bias, and it could go either way.

DR. CELLA: As far as we know.

DR. PAZDUR: Yes. That's how we would interpret the data.

DR. CELLA: Donald, you had your hand up earlier?

DR. PATRICK: I'm a little uncomfortable with the idea that effect size has anything to do with interpretation. What we're trying to do is interpret the effect size. So, I see these as just measures of distance,

and a large effect size may be meaningless -- let's hope not -- in any circumstances. But these are statistical measures that we want to put some meaning to, and it goes back to Dr. Schilsky's hypothesis-driven research because basically interpretation will depend upon our sort of theoretical underpinning about what we would consider a big change in relation to some external criterion.

Our suggestions of using global ratings of change out of Guyatt's group is, one, somewhat circular in that patients may not perceive change in certain circumstances. But we need to have specified, a priori before we go into the trial, what do we expect to benchmark our perceived instruments against. Against the patients' perceptions of change? Against another clinical outcome? And so, it would behoove anybody developing a drug to study those and think them through, prior to starting a pivotal phase III trial, to have specified what they would expect to see in the change of the external variable that will permit us to interpret the effect size that we observe.

DR. CELLA: Donald, just to clarify a possible point you're making. The usual thinking is that whereas simple statistical significance, because it's so tied into sample size, is a weak indicator of one's ability to be persuaded that it's meaningful. Effect size, because it's independent of sample size -- so, the usual thinking that

effect size, even though it's statistical, is a step up in terms of assuring one comfort that what you're dealing with is significant. Are you disagreeing with that perspective and saying that it's no better than statistical significance, or are you allowing for it to be an improvement that's not enough?

DR. PATRICK: Probably the latter.

DR. CELLA: The latter he's saying, an improvement that's not enough.

DR. PATRICK: Probably. It's a measure of distance. So, it's a standardized way of measuring the change, and there are probably -- I think I've counted 10 proposals for measures of distance from the effect size to the standardized response mean to the standard error of measurement. The papers are coming out pretty rapidly on this because it's such an important problem. But you're still stuck with interpreting the meaningfulness of that distance. So, it may help you calibrate the distance, and statistical significance, because it is sample size driven, isn't going to tell you very much.

But you're still going to have to interpret the effect size. So, it would be useful to have some agreement if this exists or it's possible that you can't do an a priori -- that this is going to be a small effect size, a medium effect size, or a large effect size, that this still

has to be theoretically driven.

Now, we may find methods -- and Guyatt's group believes that if we use 7-point response scales and 15-point global ratings of change, a change of .5, which is similar to a half a standard deviation, will cut across different therapeutic trials. I think we're looking forward to finding out that's the case, but I've never done a single study in which my external criteria all moved together.

So, it's a specification of the external criterion and some previous knowledge in phase II that may give you the idea of what are you going to do to interpret your health-related quality of life measure in the pivotal trial. This may be a clinical outcome. One of my favorites is the change in the symptom index should translate into the change in the other more distal measures.

DR. CELLA: Now, another factor is, of course, if the change occurs in a set of questions -- again getting back to the patient being the standard -- that were developed from interviews with patients so that the questions being asked have been previously endorsed by patients with this condition as being important, that's another degree of comfort that one can take, that the larger the effect size, the higher the probability that the

change is going to be meaningful. So, again, as Diane was pointing out, we're dealing with probabilities and comfort level within probabilities.

I think it's this subcommittee's challenge in the near future to put together these factors that all converge on one's comfort level that we're talking about a change that a regulating agency can consider to be meaningful based upon a collection of different pieces, and one of them may be, were these questions derived from input? What was the basis of the patient input or how was that obtained? Because the reason they may say they want that is to be able to increase their comfort level on an effect size change in that particular trial.

DR. PATRICK: That might contribute but patients may consider things important that are not responsive to change. And Lillian made a couple of very important points there, that some of these things are extremely important but will not change.

DR. CELLA: That's right. Many of us know that some of the most important things in people's lives don't change because of drug therapies, and they have to do with your social situation, your social support and your family life. Sure, I'm not saying they're not changed by disease and by treatments, but they are not usually changed differentially by drug combination A/B versus A/C or versus

placebo. So, that's an important factor.

It's another argument that Carol alluded to in terms of supporting looking at the whole picture because you would not want these things to suffer where symptom benefit might improve.

DR. TEMPLE: I'm sorry I missed the early part of the discussion. But that's been a problem with quality of life scales in all areas, not just oncology. The things that work best, like some of Guyatt's asthma scores, are fairly direct assessments of asthma, but if you then go on to ask how's your emotional state, that is, let's say, damped in comparison. One of our division directors says, your heart failure improves, you get out of your bed, and you find that the house is filthy, so you're feeling better, but your mood doesn't change.

Our response to that over the years has been to say, focus on symptoms. That's what you're most likely to do, and a perfectly good measure of whether a cancer chemotherapy is doing good is whether it improves the symptoms. So, maybe you've considered this already, but there are scales that are focused on symptoms and there are scales that are focused on the other components, the social and the psychological. Do you all have a bias about which of these is most important or, more to the point, which is most likely to be moved by an effective therapy? It seems

obvious to me what the answer is.

DR. CELLA: We did talk about earlier this morning in the first session and I have a draft statement about that, which I have to modify to be sure is comprehensive enough. But it essentially states that while it's reasonable to start from a position of looking at symptoms and focusing on symptoms as a primary an analysis, it's important to recognize that there are aspects of function that should not be expected to change and yet remain important to capture.

DR. TEMPLE: One of the things that people who carry out trials always make sure of is that the people they're looking at have or are likely to get impairment in a particular area. You know, you don't do a pain study in people who don't have pain. But the quality of life instruments we see make no attempt to get people who are particularly socially impaired or particularly psychiatrically impaired. So, how on earth can they possibly improve that? Now, maybe they could slow the rate of deterioration, but they don't try to assess the susceptibility to that. It's really a prescription for failure because the people don't have the disease they're interested in.

DR. CELLA: Your comment illustrates the wisdom perhaps or importance of being careful about planning a

primary versus a secondary endpoint. I think the perspective of the subcommittee, as I read it so far, is the primary endpoint items may be and perhaps should be identified as those things that are deemed most likely to change and most clinically relevant, assuming that they're important to patients because we're in this quality of life domain, if you will, but that there are other areas that remain important and may be superordinate over these symptoms. If they're somehow worsened, even though you're getting symptom benefit, like a pain benefit, that's a significant thing you'd want to know I would think. You wouldn't want to approve a cytotoxic that had a pain benefit but that made people so fatigued, something that's not generally captured by toxicity rating very well at all, that you didn't have the data because somebody didn't ask about fatigue. So, this is the challenge, to be sure that you're capturing enough things.

DR. PAZDUR: I think that just underscores the importance of something we were talking about as far as bringing in the investigator early on to discuss what you're going to do and also this concept of should we have a hypothesis-driven type of quality of life or symptom benefit type of analysis rather than, well, we have a colon study. Let's lop on FACT-colon on this and see what happens. Maybe these have to be not only disease-specific

but therapy-specific analyses looking at toxicity issues also.

DR. CELLA: Let me follow up on that and kind of turn back to Jeff or anyone on the subcommittee that would like to comment on this. You used the example of FACT-C. I hate to -- I don't hate it, but I'm embarrassed to use the example.

But you take a questionnaire that has a set of predefined subscales, assuming you have clinicians who are willing to take the time to look at the questions and walk through it question by question, you can go through any number of these different questionnaires, and out of a set of 30 or 40 questions, nominate a handful of symptoms that you think are not only very important -- and we know that because patients helped to create the questions, and the clinicians agree because they're helping to nominate them -- likely to change and cover the symptom map. But then you've got this problem of a handful of questions, five or six questions, that were never published in that form, never so-called "validated" as a set of questions, but were validated within a larger matrix, organized in a different way.

So, going back to your 6 out of 13, can you walk us through how you might help the investigator planning a trial who wants to be able to satisfy the FDA's

request to focus on symptoms, also wants to use a recognized, published health status quality of life questionnaire, and wants to be able to target a primary analysis endpoint that may not be previously published? What's the minimum that has to happen pretrial to be able to be comfortable with that?

DR. SLOAN: I think it follows up on something that Don was saying in particular. There is I think an idea out there that defining things in terms of effect size and so on becomes a statistical game almost or a statistical argument.

The effect size approach or the errors approach, the SEM approach, even the MCID approach for that matter, they're all statistical approaches only if the a priori work has been ignored, as you were saying. So, to use the example -- and actually I can use a concrete example where in just recently designing a trial, we are, as it turns out, using the FACT-C, because of course it was the best tool -- right? How's that for a setup, David?

DR. CELLA: Are the advertising people here?

(Laughter.)

DR. SLOAN: Put the money in the usual place?

(Laughter.)

DR. SLOAN: But realistically what we had for a particular study that's going to the North Central Cancer

Treatment Group right now, because it was a colorectal study, there were specific aspects of the disease that we thought and the treatments, I should say, that were going to impact patient quality of life in particular ways.

What we did precisely was, before the trial was started, we said, okay, there are going to be some symptomatology changes here. What are they going to be? There are going to be some quality of life changes here. What are they going to be? And listed them out, first of all, in consultation with the investigator, from the literature, from the experiences of the pharmaceutical agents that had come through phase II testing -- this was a phase III trial that we were talking about -- and basically got a laundry list of what we thought was going to change.

We then went through the FACT-C question by question and said, okay, which of these things are covered and not covered by the various items in the FACT-C. And not surprisingly, we found that many of the things that we were going to expect to see were covered by the FACT-C, and also not surprisingly, there were some things that were not covered by the FACT-C.

So, what we ultimately decided to do was to use the FACT-C in part because again there's normative data, it's an established tool, patients have been involved in the development of the process, it has been very well

delineated. And the sort of differences one could expect from the FACT-C, some of the information and data and parameter estimates, let's say, from a statistical standpoint that can be put into a power calculation to derive an expected effect size and therefore operationalize the scientific question into statistical terms so that a sample size could be estimated could then be followed.

As well, though, in particular, there was one of the agents in one of the arms that we thought neuropathy was going to be a particular problem, a particular type of neuropathy that had been observed in the phase I and phase II testing of this particular agent. The FACT-C did not have items that were specific enough, let's say -- sorry, David, but they were not specific enough to the particularly, let's say, eccentric type of neuropathy that was going to be expected to be observed in this trial.

Before you say, well, that's just a symptom, well, it is but it was a sort of subtle symptom that had come out only in anecdotal evidence in the phase I and phase II testing such that the standard CTC criterion would not have picked up anything more than a grade 1 neuropathy, but patients had told us anecdotally in the previous studies that, man, this stuff feels like bumble bees and it's just irritating as all get-out and it really impacts my quality of life.

So, what we did in that was supplement those tools with some study-specific questions derived from some wording, after having gone through the literature and pulled a couple of questions out of the literature.

And since we have no way of knowing historically how those particular instruments might behave because they were not as well developed, then we went to, okay, if we're talking about a small, moderate or large effect size, these are the sort of changes that clinically we should expect from our experience with the drug and translated that into the statistical argument in terms of effect size, and then were able to make an assessment as to whether or not our sample size that was defined for the primary endpoints in the trial would be sufficient and reasonable for the rest of it.

Perhaps that's a long-winded explanation, but that hopefully gives you an idea of the flavor of this has got to be more than a 15-minute meeting with an investigator to determine exactly what is important to the clinician and to the patient in terms of being a clinically significant change.

DR. CELLA: So, more is needed. That's one example.

Any other comments that anyone has? Dr. Williams?

DR. WILLIAMS: It seems to me that as we look into trials, we're used to looking at efficacy and also toxicity. This would, I think, really be a third domain, quality of life. Is it possible to measure or to standardize effect size in terms of tradeoff for the patient, the perceived tradeoff for a certain amount of efficacy and/or a certain amount of toxicity? Because I think that's where it's useful. Is this change worth it compared to the efficacy or the toxicity you might have.

DR. CELLA: Yes. Let me kind of rephrase that challenge in a way that also I think -- I hope -- follows up on where you were going, Jeff, and also picks up on, I think, Donald's comment and concern about only looking at effect size.

You can take a questionnaire that has a set of questions. Let's just say it's a so-called cancer-specific quality of life questionnaire. Then you decide that you need more questions, which may be a perfectly legitimate and valid decision, as Jeff just outlined, and you want to cover something like neurotoxicity. You start asking 10 questions about neurotoxicity, all the different manifestations. Yet, you don't really know how important that neurotoxicity is to the patient. You know it's important but you don't know the effect really that it has upon other areas of functioning that have more generic

importance, if you will.

If you create an index of symptom relief, benefit, plus neurotoxicity, add them together, assuming equal preferences or values on the part of the patient, then you really don't know if what you have is -- let's say at the end of the day you have more neurotoxicity measured than symptom benefit, although you did get symptom benefit. You might be in a position to say, well, the drug shouldn't be approved because there's all this neurotoxicity, but it may be because it was asked 10 times, and that went into the score.

So, I think there's risk on both sides. We want to measure things as precisely as we can, but to some extent, the more times you ask about something, unless there's some value-based adjustment or impact-based adjustment on the patient, you end up totalling up things and the score becomes a function, in part, of how many times you asked about an area.

So, there does need to be this circling back, and I'm fishing from among the subcommittee for perspective and ideas about how we can move this forward. Bob and then Rich.

DR. TEMPLE: Doesn't the fact that there's a control group help you with your concern about over-asking? That should happen in both groups.

DR. CELLA: It would help with the comparison. It would help you to believe the number, that the number is different in the treatment group or experimental group, but it wouldn't help you with knowing how important neurotoxicity is to the overall health status, well-being, and life of the patient.

DR. TEMPLE: Right. No. I see. You have to get at that in a different way.

DR. SCHILSKY: David, I don't have the answer to the question you posed, but it seemed to me that what we're addressing here is another important element that impacts on sort of this whole overall quality of life assessment. It comes back to a point that Bob Temple made earlier, which is that the more removed you get from the specific symptom complex, the more you have to consider the impact of tradeoffs. To me this becomes sort of an important confounder because you have a patient who has tumor-related pain which improves with therapy, but the therapy causes a severe peripheral neuropathy that makes it impossible for the patient to walk, and at the end of the day, the patient says, well, you know, my quality of life stinks because before all I had is pain and now I can't walk. So, has that therapy provided a benefit to the patient or not? Those kinds of tradeoffs I think become very important in these analyses.

DR. CELLA: Stacy.

DR. NERENSTONE: Getting back to what Dr. Williams I think is asking, I think it gets even more complicated because a certain symptom complex is going to be recognized differently by different patients. And you may have a 90-year-old patient who says, absolutely not, I don't want that treatment with those resulting disabilities, and you may have a 45-year-old patient who says, absolutely, if you can tell me there's an X percentage chance that my tumor is going to shrink in my liver, I will tolerate being in a wheelchair. So, I think it's very difficult for us to sit here and say there is a percentage of disability or based on quality of life to vote a drug down.

DR. PAZDUR: I think that's very important. You've been talking about symptom toxicity, but the other part of that triad is efficacy survival, and that needs to be factored into here. Obviously, whenever we make a decision on a drug, all three components come into play here. We can't just isolate symptoms, quality of life, and sometimes the primary endpoint of the study and the primary reason why we're giving the drug. Obviously, they all are interdependent somewhat on each other when we're making this decision as far as drug approvability.

DR. CELLA: Carol?

DR. MOINPOUR: Well, related to what you just said, I was just going to emphasize that ODAC, clinicians, medical oncologists have been dealing with multiple endpoints for a long time, so the additional information and maybe some conflicting information in quality of life data is just another piece of that puzzle that people have had to present to patients on the pluses and minuses of a particular treatment. I think it helps. It gives more information.

DR. CELLA: Bob?

DR. TEMPLE: I think the point Stacy was making is extremely important, that the different events, the benefits, risks, and one's judgment of them, are different for every person. This is partly a lumpers/splitters' argument. Our inclination historically has been to try to define the bad things, try to define the good things, and let individuals and their caregivers work it out.

In some ways, a global score defeats that a little bit or is an opposite view, which says you really need to look at the net for a large of group of people, but it masks the fact that attitudinal sets and preferences and what happens to individuals could lead to very different outcomes if you look at particular people.

DR. CELLA: Jeff.

DR. SLOAN: To follow up on what Carol said,

oftentimes in designing trials for efficacy, we talk about, for example, in lung cancer will this agent improve median survival by 3 months and go back and forth about whether a 3-month improvement in median survival is actually worthwhile. Is that clinically significant?

I think the arguments are no different really for quality of life. It's just a different aspect, especially when, going back to the lung trial example, okay, maybe the new treatment, for example, hypofractionated radiotherapy, can give potential benefit of around a month or 2 months median survival with an incredible increase in associated toxicity. As you said, the 84-year-old perhaps will say, well, maybe that's not worth all the trouble, but another 84-year-old might say, but, you know, my granddaughter is getting married next month, so I'll go for that extra month. I think in terms of quality of life issues, the arguments are the same.

Again, this idea of going through an individual index of like one score for quality of life I think is a bit of a -- going back to the blood pressure example that I gave, I'm not sure that's achievable. In the same way that we present, as you said, treatment trial results, well, X percent of patients have a certain probability of a certain degree of survival benefit, there's some toxicity, I don't think it's unreasonable to add to that, but a certain

proportion of patients experience improvements in quality of life in this way, shape, and fashion. It's going to be a multidimensional argument. I don't think there's any way of getting around that.

I think it's fair to say your mood might improve. Your physical functioning may be decreased a little bit. You may feel like withdrawing. I think that's where you get into the interaction with the patient and the clinician presenting all the data that's available to the patient so they can make an informed decision.

DR. CELLA: We're about 10 minutes from our next break, our closing of the session, for lunch. Nobody has mentioned preference-based measures yet. So, I bring them up, first of all, to put them on the table because I think they should be there to fill out the discussion, but also as a possible way to provide perhaps not the most sensitive to change assistance, but some kind of overview, global if you will, aggregate sense of the value of the health state that's the ultimate bottom line for the patient.

I think there has been a hesitation to use these instruments in trials because they're difficult to administer in part, but as they become easier to administer, there's a concern about sensitivity to change that's the next sort of fear-related component.

But if we, for example, said they were not inserted to be expected to change over time or even necessarily to provide a denominator for a cost effectiveness analysis, but to be able to get a sense of the full picture, how would you -- since we're not here to talk about cost effective analysis and qualities and any of that -- but it strikes me that this could be something that could be recommended to be inserted for a very different reason, in a sense, which is just to make sure on a global basis that you're not making too much out of neurotoxicity or some other side effect or some benefit that you're imparting.

Diane.

DR. FAIRCLOUGH: David, could you just be really clear what you mean by preference-based because I think there may be some variation or lack of understanding what that is exactly.

DR. NERENSTONE: Or even define it for the non-QOLers among us.

DR. CELLA: Sorry. I apologize.

A preference-based measure, as opposed to a health status measure, is one that, because of its grounding, the way it was developed, using input from community populations or using theory, utility theory, generates a score between 0 and 1, where 0 is meant to

represent a health state that one values equivalent to death and 1 is meant to represent a health state that one values as equivalent to perfect health. The score ranges anywhere between 0 and 1 and is then typically used to modify survival time in a quality-adjusted life analysis.

What I was saying was that a different possible value of such a number could be to run a check, if you will, on the approach that we're kind of driving at here which is to load up your trial with disease-related symptoms and side effects and then make a conclusion. How do you make a conclusion if you don't know that you're really capturing everything? And this may be a way to at least say, well, you know, there's this kind of benefit conferred by the treatment, there's this kind of toxicity conferred by the treatment, the numbers seem equivalent.

We don't know if they're valued the same by people because we can't tell that from the questions themselves. As I mentioned, the number you get is in part a function of how many times you ask about it. But we have this other number from 0 to 1 that was generated to at least see if there's not some major disconnect between the detailed, more sensitive data and this broader value-based number.

Was that clear? Stacy, you need more?

DR. PAZDUR: Do you want to give us a specific

example maybe?

DR. CELLA: Okay. We'll go with the neurotoxicity example. Let's say that you have a drug that doesn't change survival, might have a modest benefit to progression-free survival, and tied to that benefit to progression-free survival you have good indication that there's symptomatic benefit that seems to be related to drug effect, at least partly related to drug effect by virtue of its effect size, and the set of questions that were pulled together to measure pain and whatever else might be associated with tumor-related symptoms, whatever tumor-related symptoms there are. There's also toxicity in the form of, say, fatigue and neurotoxicity. That was measured because the trial has put in fatigue and neurotoxicity questions.

And in the end, you have this picture where you've got pain benefit, relief benefit, a little progression-free interval benefit, added fatigue, added neurotoxicity. It all seems to be there and you're not sure if you should approve the drug.

Well, then you might look to this number between 0 and 1 and ask is there any indication that the overall value that these patients placed on their health or that another group of people would place on the health states described -- it doesn't have to be collected from

these patients -- that it's different. And that may help balance the scale.

DR. FAIRCLOUGH: David, you might give an example of how you would get that number between 0 and 1.

DR. CELLA: Well, you can get it directly from the patients, which is controversial because patients have a stake and a bias in reporting their condition and actually tend to report higher numbers, report their health as better. Or you can get it from a representative sample.

You mean how you get it?

DR. FAIRCLOUGH: How you get it.

DR. CELLA: Okay. I'm trying not to take too much time with this.

There are several ways to get it. The original way is to present a gamble to an individual and say, imagine that there was some risk of death that you could incur, but in exchange for that risk, there's a different risk for perfect health. You basically find out how much this person is willing to risk death in order to achieve perfect health, and the more risk they're willing to take, the worse their health is likely to be because they're telling you that they're willing to take a bigger risk.

Another way is to see how much time people are willing to trade to get perfect health.

Still other ways are to administer health

status appearing questionnaires, but because the states described by the questionnaire have been anchored to community populations, that number is then derived from the score that the person gives you.

And I'm either digging you deeper into a hole or helping to clarify. Stacy.

DR. NERENSTONE: I'm not sure I understand. Are you looking at a change in that number with time, or are you looking at a number at some prescribed point? You're looking at a change, pretreatment versus post-treatment.

DR. CELLA: Right.

DR. NERENSTONE: I guess that gets back to however you derive it, it gets back to Dr. Temple's question or concern that by lumping there are so many confounding factors, especially in a phase II study where you're just looking to see does the drug have activity. The vast majority of patients are going to progress. The huge number of patients are going to progress. So, I can probably guarantee that the score is going to be worse for the great majority of patients at the end than at the beginning. And by lumping it together, I think you're going to obscure any differences rather than show any differences.

DR. CELLA: Well, I wouldn't suggest this and

didn't mean to be suggesting this as a replacement for looking at the things that are likely to be more changeable, more variable over time as a function of the treatment. But to the extent that the effects of a drug go both ways -- you know, if everything gets better or everything gets worse or there's no evidence that anything changes one way or the other, the decision is pretty easy. The decision is not as easy if you're looking at data -- I'm trying to put myself in your situation, looking at data where the disease-related symptoms seem to be improved. There may be some benefit to progression-free interval. There's no overall survival benefit and you're aware that there's toxicity. How do you then decide this is a drug worth approving?

Well, one thing that can help you in that situation is to look and see if the value for the health states of one group versus another are indeed different, and if they're not, then it might support a view that it's a wash.

The reason I brought this up is because once you start moving in to tailoring your assessment and extracting questions and creating new indexes, which is a compelling thing to do, you run the risk of stacking the deck one way or another in favor of a treatment, if you load in questions about the benefits you expect, or making

the treatment look bad, if you load in questions about toxicity. So, that's going to be sort of a risk out there in every trial that proceeds to do this, and the question was how do we deal with that.

Well, there are ways.

(Laughter.)

DR. CELLA: And I'm sure this group is aware of them.

I think our task is to pull together enough of a consensus, if you will, of acceptable approaches. Again, in this area, like the first area where we talked about we're not going to, at the end of this process, however long it takes, be recommending one questionnaire or one measure. We're also probably not going to recommend one approach to clinical interpretation and clinical significance. Our task is to outline the considerations that are critical in planning a trial so, at the end of the day, there's some acceptable data, recognizing that the field is moving and improving and that this will need to be a set of recommendations that moves and improves with the field. Reasonable?

So, we'll break for lunch. For the last hour, I'll try to come up with some summary points from this session, along with the others, to bring back to you for our course after that. Let's take a break and we'll see

you again at 1 o'clock.

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

AFTERNOON SESSION

(1:08 p.m.)

DR. CELLA: Welcome back. If the committee members could please sit down and we can get started again with the afternoon session.

I have a couple of brief comments, announcements. One is a small change in the agenda. One thing that's not changing is the time of adjournment. We will definitely adjourn at 4 o'clock. So, those of you who are concerned about flights, we will be finished at 4:00.

The second thing is we're going to begin at 3 o'clock with some input from the FDA on specifically what they're looking for from us, in part driven by what they've heard so far. So, we want to have the opportunity to get a refocusing, if we need it, from the FDA. And I've asked Dr. Beitz to either do that herself or ask Dr. Pazdur or Dr. Temple to do it. So, one of the three of you I'm hoping will say it. Let us know at 3 o'clock where you are with this and what you're really looking for so that we can then plan for that last hour for the next meeting.

Now we move to the open public hearing section again, and we'll start with Dr. Rick Berzon. Rick?

DR. BERZON: Thank you, Dr. Cella.

I'm enjoying this quite a bit. I'm Rick Berzon. I'm with Boehringer-Ingelheim.

DR. CELLA: Excuse me, Rick. I'm being asked to ask you to come to the podium please. Thanks. That way we get you on film.

DR. BERZON: I'm Rick Berzon. I'm with Boehringer-Ingelheim Pharmaceutical Company. My background: I'm a doctor of public health, and my background is in epidemiology, health services research and clinical medicine.

I don't have a prepared statement. I just wanted to say that this is a subcommittee whose time is overdue, and I'm delighted to see it here and ongoing.

I think many of us who work in industry are occasionally confused, if not uncertain, as to the kinds of endpoints to put into trials so that we can both address the regulatory requirements, as we understand them, and so that we can promote quality of life. What I mean by that is that it's not always clear to us exactly what kind of information is acceptable.

Perhaps my remarks are premature. I wasn't aware that FDA would respond specifically to this issue, so I applaud and I look forward to hearing it.

But there's often confusion with respect to measures, the extent to which a measure has to be psychometrically sound and what does that mean to the FDA. If we do two trials and we use two different measures and

they don't necessarily demonstrate the same effect, how do we interpret that? Issues around sample size, which I understand that we're still in the process of discussing many of these issues. To the extent that we could get clarity or at least direction on some of these points, it would aid us enormously as we attempt to design studies that can truly measure quality of life -- that is, my understanding, subjective health status on the part of the patient -- and whether or not that needs to include symptoms.

Oftentimes when these measures are developed and we go directly to patients to develop them, patients don't necessarily differentiate between symptoms and what they perceive to be quality of life, and I think this point was made earlier by Dr. Sloan.

But if we could get some guidance on this so that we could better design studies that will benefit us as an industry and the people for whom we develop medicines, that would be terrific.

Thank you very much.

DR. CELLA: Thank you, Rick.

Susan Weiner from The Children's Cause, Inc.

MS. WEINER: I'm Susan Weiner. I was originally trained as a developmental psychologist and was the parent of a child with a brain tumor for more than 13

years. I'm President and founder of The Children's Cause, which is an advocacy and education nonprofit, dedicated to pediatric cancer issues.

My message here today is quite simple, which is don't forget the kids. It's a message that derives really from both of my experiences, that is to say, don't forget the measurements of the quality of life of kids depending on their developmental stage, and also don't forget that survival as an endpoint has very different meanings for kids depending on how old they are.

Finally, I think that paying attention to quality of life of pediatric brain tumor patients which is the most common solid tumor these days in kids and really the next frontier, and hopefully one of the last frontiers in pediatric cancer, the quality of life of the kids in the trials and the parents' experience and need to protect the quality of life of the kids in the trials is a very important consideration in designing them.

Thank you.

DR. CELLA: Thank you. Much of what we discuss and decide and recommend should apply comparably to children as adults, but there certainly are development-specific issues so need to be considered. So, thank you for the reminder and for the call.

Leonard Rosen, Cure for Lymphoma.

MR. ROSEN: I just wanted to briefly comment on some of the things that were said in the discussion.

I'm an indolent NHL, having been diagnosed two years ago. I've had no treatment whatsoever, nor is there any clinical trial I think that I could be treated in. I'm one of the 0 to 1 perhaps that doesn't have a clinical trial.

I just wanted to say that I think I applaud the purpose of this meeting and the idea of embodying quality of life to a further extent in the approval of drugs. I think the effort to standardize the process and to standardize the format perhaps to some degree is worthwhile.

The things I want to express was the caution that cancers are unique and we're learning that it's not 10 diseases or 100 diseases, but perhaps 1,000 diseases. Individuals are unique and they're infinitely different.

Accordingly, I think we should not be too rigid in the creation of the formulation, particularly the first time you do this. You ought to leave flexibility. I'm a lawyer by profession. Leave some rubber so that you can, in fact, develop ultimately perhaps a format that is more specific. But going into a process like this, I think you have to leave room for things to develop.

I think it's easy to applaud measurement, but

the idea of creating a formula by which you arbitrarily measure and then say, well, based on this measurement, I'm going to do X or Y, it just seems to me is a foolish objective. I'm not saying that you think of that objective. But there are always questions about science and there are questions about the quality of life criteria, but there are also questions about the so-called scientific parts of what the committee hears when they're approving drugs. I think you want to know as much as you can know about these things, but the ultimate decisions require discretion.

There are many factors to be considered, and I don't think you ought to do anything that short changes the need for discretion, the ability to use discretion in deciding whether to approve a particular thing. Measuring quality of life versus survivability versus the efficacy of the drug, the toxicity, all of those things go into it. It may be a slight difference in quality of life may influence a decision and a great one may not influence a decision in a particular case because of other circumstances.

It's a very complex process, and I just hope you keep that in mind as you do this and don't create something that's too rigid.

Thank you.

DR. CELLA: That's a good caution. Thank you

very much.

The good news is that we really may be -- even if we wanted to create something rigid, we may be forced to keep it open enough to satisfy everybody. So, I think the outlook is good that this will be flexible. We do need to make it specific enough so that there are good guidelines.

There are two people from this morning that were not here this morning, might be here this afternoon: Jan Maryak or Nancy Roach. Just checking to see if either of you is here.

(No response.)

DR. CELLA: Is there anyone else that would like to say anything?

(No response.)

DR. CELLA: Okay. We're okay. Thank you.

So, we move to the next part of the agenda which requires a phone call be made, and we'll get Dr. Nan Laird, who is at the Harvard School of Public Health in the Department of Biostatistics. While that call is being made, let me introduce two biostatisticians from the FDA or let you introduce yourselves. Claire?

DR. GNECCO: Thank you, David. Claire Gnecco, Center for Biologics, Division of Biostatistics.

DR. CHEN: Gang Chen, Biometrics, CDER, FDA.

DR. CELLA: Thank you. We're getting Dr. Laird

on the phone, and maybe while we're doing that, Diane, if you want to step up and prepare yourself and your slides. Are you using the LCD projector?

DR. FAIRCLOUGH: Yes.

DR. CELLA: Hi, Dr. Laird. Hello. This is David Cella from the Quality of life Subcommittee meeting at the Oncologic Drugs Advisory Committee meeting at the FDA. Are you able to hear us okay?

DR. LAIRD: Yes, I can hear you fine. Thank you.

DR. CELLA: And can the audience hear Dr. Laird? Raise your hand if you cannot. I think we've got the mike on you and it's all working. Congratulations. These are the kinds of things that usually don't work out, and it's wonderful that it did. Thank you. Thanks to Dr. Somers.

DR. LAIRD: Actually I have them right in front of me.

DR. CELLA: She's got copies of the slides. Okay. We're going to proceed, Dr. Laird, with Diane Fairclough's presentation, and then we'll look forward to your comments.

DR. FAIRCLOUGH: When I talk about quality of life in oncology, this is one of my first slides usually, sometimes the label of my talk. I feel like I had some

plans in the audience to make this first point.

Too often exactly what has been said a couple of times is that a quality of life assessment has been added to a clinical trial without a lot of thought about what is the question and what we're going to do with the data. So, before we jump into issues of missing data and summary measures and longitudinal studies, I think it always has to be in the context of what is the question, and unless we know what that question is, we're not going to be able to decide what's the best strategy.

In most trials, we have a univariate outcome. Survival is a univariate outcome. Disease progression is generally a univariate outcome. And so, when we say we're going to look at whether treatment A has a better survival than treatment B, we're really clear about what the question is. When we say does the quality of life of a patient in treatment A differ from the quality of life in treatment B, we haven't defined the question at all, and part of that is because we have a multidimensional construct, but it's also because it's something that's measured over time.

Some of the things that we need to define, before we try to handle how we're going to analyze or even how we're going to design the study appropriately, is what is the objective. Are we looking at comparisons between

treatments? Are we looking for a change within a group over time? What is the population that we want to do our inference on? Is it all patients that were randomized to that study? Is it only while they're on treatment we're going to look at their quality of life, or are we trying to look at some of the issues of their quality of life as survivors when they go off treatment?

We have to think about the time frame. All these things have to be defined.

So, only when we have clear and specific objectives can we define the design and the quality of life. Do we know how long and how often to assess the quality of life? Do we know what type of measures that we want to put into our assessment? What may be appropriate for a patient on treatment is not going to be an appropriate measure for a survivor because they're going to have different issues.

So, let's assume that we actually have a well-defined question. Then as we're starting to think about our analysis, one of the big issues is missing data. And why is it a problem? There's a minor problem in the loss of power to detect differences. That's something we can actually fix by increasing our sample size, but the major problem is that there's a potential for bias if that missing data is related to the individual's quality of life

who we don't actually measure the quality of life on. So, that may be affect our treatment comparisons and the inferences. I talk about this in the context of quality of life, but it's the same issue if we're measuring pain or any other outcome that might be related to the response.

One of the typical questions that I get is people want a very simple answer on how much missing data should be allowed. Unfortunately, there is no magic rule. It really depends on the setting and the research question and what we're trying to do with our inferences. What would be acceptable in an adjuvant breast cancer study would be very different than what would be acceptable in a pancreatic cancer study because we have a real difference in the mortality and the morbidity of those patients and our ability to follow up.

And it may be very different depending on what our question is. Are we talking about claims of improving quality of life or are we making comparisons between treatments? So, it's conditional on the patient surviving possibly. So, setting one rule is just not going to work for us.

The type of missing data and why it's missing is very critical to any assessment of whether it's a problem or not. There are three classical definitions of types of missing data.

The first is what we call missing completely at random. It's going to be that the patients didn't get their assessment because there was a snowstorm and they couldn't get to the clinic and clearly unrelated to their quality of life. This is going to be very, very rare in oncology trials. In most cases it's going to be that the missingness of the quality of life assessments over time are going to be related to both the quality of life of that individual previously and their current quality of life. So, anytime that patients with poorer quality of life, for example, at baseline are more likely to drop out and it's very predictive, then we can't make this missing completely at random assumption.

Missing at random allows the dropout, with respect to quality of life, to be dependent on previous quality of life assessments. So, it might be the quality of life that the previous measure predicts, whether the patient will have a missing assessment or not. This is definitely more likely than the missing completely at random.

But in the oncology setting, we're actually more likely to have the setting where the reason that the assessment is missing is related to the actual value of the quality of life of that person at that time. So, patients currently experiencing more toxicity are more likely to be

missing their assessment than people that aren't. So, especially in the advanced cancer, we're going to be in the situation of missing not at random.

So, why does this matter?

The next thing is can we test for these different things. We can test the difference between missing completely at random and missing at random because we can actually set up a model and test whether the missingness depends, for example, on the quality of life at the previous assessment. But what we can't test and what's very problematic is we can't test between missing at random formally and missing not at random. The reason is because the information that we need is what we're missing to do a formal test.

However, when we have other clinical outcomes and measures, death, toxicity, disease progression, symptomatic disease progression, and we know that those are related to the proportion of missing data we're observing, we can't dismiss the fact that we probably have not missing at random data. So, while it's not a formal test, it's something that we need to look at, and when we see this type of pattern, we have to consider the possibility that we have non-ignorable missing data.

Well, just to talk a little bit about why this is important is because different methods of analysis make

different assumptions, and if the assumptions aren't met, then there's a probability that there's going to be some bias in the estimates we obtain. The methods that assume that it's missing completely at random are ones that are often used and often presented, and I don't know that people always understand that they're making that assumption.

Things like MANOVA, which excludes all patients who have any missing data, consistently over every quality of life study I've ever looked at the data, the patients who are completers have better quality of life than the non-completers. That's even true if you take a group of adjuvant breast cancer patients, all of which are disease free. Even given that you don't have evidence of progressive disease within that group of survivors, their quality of life is related to the missingness in the data.

Another typical analysis is actually to do repeated univariate t-tests. The problem is it totally ignores any information, for example, from the previous assessment. When you compare the second assessment and third assessment, you ignore all the data at the other assessments. That's one of the assumptions that you're making in there.

My feeling is we should never be using these types of analyses that make this restrictive assumption in

the analysis. One of the reasons is because, at the minimum, we have good analytic methods that are easily accessible that make the less restrictive assumption of missing at random. Now, this may not be enough, but it may be reasonable in settings where we have a very small amount of missing data where there's minimum morbidity and mortality and for certain restricted questions. And mixed effects models and repeated measures for incomplete data, PROC mixed, are methods that we can use to do this data analysis.

The real challenge methodologically is that in many cases we're looking at settings where we have the data non-randomly missing. Unfortunately, there is no way to say this is exactly the right method because what we need to really test whether it's the right method is exactly the data that we're missing. So, all the models are somewhat untestable in their validity.

But we can look at various models under various assumptions and we can get a good sensitivity analysis and see whether our results are consistent under different assumptions. And that's just how we have to go in this setting.

What's so critical is to understand the assumptions under these methods, to make sure that you understand what you're doing and also have good clinical

correlates that help you with these type of analyses.

This is an observation. A lot of times I'm asked, well, we have the same pattern across both arms. Can we then ignore it? I'm really uncomfortable with saying we can ignore it. However, so far I don't have a good counter-example, and that's what this last point is saying. It's not that I'm advocating ignoring it when we have exactly the same missing data patterns. It's that often treatment comparisons -- what happens is the bias is consistent across the two treatment arms, so that when we take the difference, that bias difference disappears. But it's not a guarantee. There's nothing that guarantees that.

So, you've just gotten in 5, 10 minutes what I usually take 2 days to discuss in terms of missing data. It's obviously a complex problem.

But the other issue that we have in analyzing and interpreting is the multiple endpoint issue, and it comes from multiple domains and from longitudinal assessments. It creates a major concern about the multiple testing issue, as well as interpretation of so many sets of p values.

So, what are the possible solutions? Well, one suggestion is often to limit the number of primary hypotheses, but then somebody would say, well, why did you

collect all the rest of the data? Often descriptive statistics are done, whether they're in terms of plots or just estimates or estimates with confidence intervals. Actually there's implied testing there and we're just kind of avoiding the problem.

Another set of strategies are alpha adjustments and closed testing procedures. Probably the least desirable of all these is to do a Bonferroni correction, dividing by the number of assessments times the number of domains. That's when you get into the power problems.

Then the third option is to use summary measures. A summary measure might be the area under the quality of life versus time curve, or it might be a time to an event. What's really going to be effective and probably most useful is to use some combination of all of these.

Just quickly some of the advantages and disadvantages, limiting the number of primary hypotheses, the alpha adjustments. There are closed testing procedures. You reduce type I errors, but you have some loss of power. You still have the large number of tests.

Summary measures. You can increase the power to detect small, consistent differences over time. So, you may not have a huge impact of quality of life at any one time, but if the quality of life of a certain group of patients is consistently better, then that is probably more

clinically relevant. Probably one of the best parts is you have fewer tests to interpret.

The real critical thing is picking the right summary measure. You need to have a perspective on, again, what is the question. An AUC wouldn't be necessarily the best measure for trying to look at the delay in the onset of symptoms. You might then use the time to some change. So, you really have to relate it back to the expected pattern of change in that population with that drug and what's the question. So, good summary measures really help. Bad summary measures just make things disappear or it's confusing. There's too many tests.

So, my summary is, unfortunately, one size is not going to fit all. I can't give you a nice, easy formula for handling the analyses of quality of life. But careful planning in design phases is so critical, and you can minimize a lot of problems by first getting a well-defined objective, but then also thinking about strategies for minimizing missing data, answering questions about do we want to get assessments after a person has relapsed and gone off the drug, is that relevant to the question.

Just kind of a comment. One of the strategies that seems to be being used a lot is that when quality of life is a secondary endpoint, people often delay writing the analysis plan. There may be some benefits, but there's

some down side. One is that they just kind of push the problem back. So, we get back to my first slide: Now that we have the data, what do we do?

There may be some advantages to delaying some fine decisions to having a blinded look at the missing data patterns and the proportions may help you understand the choice between two possible strategies. But I don't think that it should be delayed completely, and you should have some thoughts about what you're likely to see and how that might affect your analysis plan. Otherwise you're just going to find out that you're stuck. You didn't think about something because you didn't define it.

DR. CELLA: Thank you, Diane.

Dr. Laird, did that come through okay for you?

DR. LAIRD: Well, first I want to thank Diane for doing a really very nice job. (Audio interruption) agree more with her first few slides that you really have to decide what it is you want to measure because (audio interruption) but (audio interruption) some very basic (audio interruption).

DR. CELLA: Dr. Laird, I hate to interrupt you, but you're coming in and out, and I wonder if it has to do with some feedback in the equipment here.

DR. LAIRD: Hold on a minute.

DR. CELLA: Sometimes that happens with speaker

phones when you've got some -- let's try turning some things off, maybe turn the mikes down. Why don't we turn some things off? We'll call her back.

(Pause.)

DR. CELLA: Could you just start from the top, Nan?

DR. LAIRD: Yes.

DR. CELLA: Thanks.

DR. LAIRD: I wanted to say that I couldn't agree more with the beginning questions that Diane laid out.

I'm hearing some feedback now. It sounds like somebody is hammering. Do you hear that?

DR. CELLA: No, not on this end.

DR. LAIRD: Okay.

DR. CELLA: Is that better?

DR. LAIRD: No. Every time I say something, I hear a hammering noise, but if you don't hear it, then we'll go ahead.

For example, Diane laid out the issue of population right in the second slide. Should we be talking about an intention to treat type analysis? Should we restrict ourselves to quality of life of patients who are still alive or responding on therapy, or should we restrict ourselves to patients remaining on therapy?

Even if you make those basic distinctions, there are still additional questions as to how one might handle missing data. For example, if we take a sort of intention to treat analysis, but a substantial number of patients dropped out and no quality of life measurements are available on those people after they've dropped out, then all of the methods for getting a summary measure of quality of life comparing the treatments effectively are making some assumptions about what's happening to the quality of life measured after patients drop off the study, so that you could envision answering questions like what is the quality of life experience of all patients randomized to this trial, assuming that after dropout their quality of life trajectory looks the same as people who didn't drop out. Or should you make some other assumptions about the quality of life for patients who have dropped out and on whom you have no additional measurement?

If you choose only to look at quality of life among patients who are continuing on the therapy and who are responding to the therapy and who get measurements on quality of life, well, that's a different sort of question.

These are the kinds of questions that I think need to be discussed between statisticians and clinicians and other interested parties as to what we really want to try and measure. So, I think Diane's summary there of what

are the clear and specific objectives, if you have them clear and specific in the beginning, then they will really define how you do your design and how you set up your analyses, and they solve a number of your analysis problems.

Now we get down to this. Of course, one of your big problems is missing data.

But let me stop for just a moment. I don't actually know what the typical protocol is in the typical cancer study, but I know that in some studies, once patients are removed from a particular treatment protocol, then additional measurements are not made of things like quality of life or additional kinds of assessments that might be made.

I don't know what the situation is in cancer, but I do think that one of the things you need to do is define a clear period of time in your protocol -- say it's two years, three years, whatever it is -- and regardless of whether or not the patients stay on or off the therapy, one should continue to get the quality of life measurements throughout the duration of the study.

Sometimes this is not so desirable because patients may be very ill and it may be viewed as too much of a burden to require patients who have withdrawn from a protocol to continue to make these kinds of measurements.

But you could always consider at least making a few number of measurements after patients go off the protocol, randomly select a few patients to continue with the measurement schedule, or make a fewer number of measurements after patients are off the protocol.

So, I think one should take the strategy of minimizing the number of missing measurements that are there due to the design of the study, because often missing measurements are designed into the study by saying once the patient goes off the protocol, they don't have to have any more measurements. That kind of thing I think can be avoided and that can help in a number of analytical problems.

If you're interested, for example, in the kind of intention to treat, what is the quality of life of patients randomized to this therapy over the entire duration of interest, regardless of whether or not they stayed on the therapy, they went off the therapy, they went on some other therapy, you can only really do that by continuing to take the measurements after they're off the protocol.

But, of course, you're going to have missing data. In general, I agree with Diane's point that there is no single rule as to how much missing data is permissible, although in practice I tend to find that 5 percent is not a

bad rule just because I find that in the kinds of analyses that I've looked at, having 5 percent or less doesn't actually make a big difference in terms of the results.

Diane then elaborated the various levels that statisticians use for describing missing data. It has always seemed to me that quality of life is a very clean example of what Diane was referring to as missing not at random, and that is that a patient's decision to fill out the quality of life endpoint probably depends rather more on what their current quality of life is than what their previous quality of life is. So, it depends upon the value which you may not, in fact, observe.

Diane also makes the important point that there might be many other events that are somewhat ancillary to the particular question at hand, although nonetheless extremely important, clinical events like toxicity, disease progression, et cetera, that might, in fact, predict quality of life and they might also predict whether or not patients respond at that point in time. So, that's another thing to keep in mind that you might want to get as much information as you can about these ancillary factors that would affect both quality of life and the likelihood that patients respond. Some of those may, in fact, become quite useful in terms of trying to do an analysis that teases out the effect of missing data on the results having to do with

quality of life.

Diane did point out that we now have a number of techniques, and the one which I think comes most defined is using the PROC mixed and SAS now has the ability to do a maximum likelihood analysis of repeated measures type data when you have missingness that does a type of missing at random. But I think what is important to remember about these analyses is that the analysis can only utilize the data that you give it. So, if you're only giving it information about which treatment a patient is on and the repeated measures of quality of life, then these other variables, these ancillary variables, which may be quite important, aren't being taken into account.

With regard to more complicated types of analyses, there are several types of analyses, selection modeling and pattern mixture modeling, that people have developed for missing not at random data. Like Diane mentioned, these provide people with a way of doing sensitivity analyses because, in order to do them, you have to make a fair number of assumptions about the distribution of the data and the missingness process as well which, in fact, in general are not testable in your data. So, I think of them as just providing people with a way of looking at some alternatives to the standard missing at random answer that you get that may display certain

sensitivities to missingness that is not at random.

These analyses, though, aren't easy to implement. They're not off the shelf. They don't have standard computer packages that people could use to implement them. So, I think that if we really want to start requiring that people do them on a wholesale level, then you're going to have to get statisticians to work much harder to come together with a consensus on how these different types of missing not at random analyses should be done.

Then I'll make a last comment about multiple imputation methods. Multiple imputation methods are another type of approach that in many ways are not dissimilar from the PROC mixed type approach, and they have been advocated by a number of people. In fact, there is now a sort of first generation of commercial software doing multiple imputation analysis of repeated measures data with nonresponse. But I think of it as very much a first generation software and it has a lot of features about it that aren't really, I think, appropriate for this setting.

But I mention it because I think multiple imputation does have sort of an advantage to go a step beyond what PROC mixed does without taking the full step towards the missing not at random, and that is the central way that multiple imputations work is very intuitive. You

impute values that are missing. The advantage of multiple imputation is that it can allow you to impute the missing values. You can allow those to depend upon these other types of characteristics such as presence of toxicity or disease progression and so on and so forth. So, as long as you do have individuals in the study who have quality of life measures under those adverse conditions, then you can make imputations for people who are missing the quality of life measurements who also have those same adverse conditions. So, it allows you, in a little more natural way than the PROC mixed framework, to include that additional information.

But as I say, even with multiple imputation, it's not a technique that can be done automatically, routinely, and it's not computer automated right at this present time.

So, there again, if we're talking about doing analyses which really try and do an honest assessment of this complicated problem of how to deal with missing values, which may be due to the design of the study or which may be due to conceptual issues -- an example that I have of the conceptual issues is, does it make any sense at all to impute missing values for people who have died because of the disease? This has always been a very big stumbling block for me in terms of doing these kinds of

repeated measures when the missingness arises because the patient has died because of the condition under study. Then what sense does it make to actually think about in some way measuring the quality of life or imputing the quality of life for that patient after death. I think that's one of the kinds of questions that this group ought to try and address.

So, that's really all I have to say.

DR. FAIRCLOUGH: Nan, you just dug us a deep hole.

(Laughter.)

DR. CELLA: Oh, boy. You've offered an awful lot. Thank you very much, Nan.

DR. LAIRD: Okay.

DR. CELLA: You're going to stay on for a while till you have to go teach this to some students.

DR. LAIRD: Sure.

DR. CELLA: Dr. Laird needs to leave us at 2:30 to do some teaching, probably a little before then. But thank you for staying on for the next 20 minutes or so.

DR. LAIRD: Sure.

Could I ask that people in the audience speak up because I actually had kind of a hard time hearing Diane.

DR. CELLA: She needs to be at a microphone.

We'll be sure to do that.

Okay. Who would like to start?

(No response.)

DR. CELLA: Well, I'll start with something that is among the many things that you raise. I think, at least of the five main points I picked up, it was the second one, and it has to do with the recommendation to continue to gather quality of life data even after the patient switches off of the study drug or is taken off the trial, if you will, but is still alive. I think we all understood the case that you made for that, but let me offer -- and this is not my personal opinion, but it's what I understand to be, if you will, the opposing view that has I think often been raised by statisticians themselves.

That is, when the patient comes off study, he or she will switch to another treatment that could then be the causal agent for the change. So, when you're evaluating the change that occurs after the patient switches to a new drug and attributing that in an intent to treat fashion to the experimental drug, there's a problem of interpretation and attribution that the trialists would like to avoid.

Could you comment on that?

DR. LAIRD: Yes. I agree that that's definitely a problem. What does typically happen in these

trials? Does the patient go on a standard therapy? Are they put on a different therapy under study? Or what typically happens?

DR. CELLA: It varies. Dr. Pazdur?

DR. PAZDUR: Yes. It varies tremendously from study to study. So, post hoc assessment of patients after they come off trial is very difficult. At best we could get survival data, but trying to ensure subsequent therapies and mandating subsequent therapies usually cannot be done in the context of even the most sophisticated trials.

DR. LAIRD: Well, but let me ask you this. If you were in a study and the final endpoint was death within 3 years and somebody drops off after 6 weeks, and you don't really know what happens to them, don't you in an intention to treat analysis still follow that person for death?

DR. PAZDUR: Yes, we try to do that, of course.

DR. LAIRD: In the primary analysis they're included in the treatment to which they were originally assigned, even though they might possibly have spent the vast majority of the 3 years on a different treatment.

DR. CELLA: Let's assume that the logistic problem, the practical problems are not the factor; that is, if the data were important to collect, the trialists would find a way to collect the data.

There still remains the issue of this interpretation where I think some people would still reasonably argue that they would rather not collect that information because they wouldn't know how to interpret it.

Diane?

DR. FAIRCLOUGH: I can't address your concern completely about the heterogeneity of the additional therapy that somebody goes on. What would be an issue in a headache trial is really, I think, different than the issues that we have in oncology. My concern is by not following the patients after they have shown radiologic progression is that I think probably when we're going to see some of the biggest differences that are associated with the disease or the failure of the drug to control the disease is going to be as the patient moves beyond radiological progression to symptomatic progression. Actually the period just prior to death is really when you see a lot of change in the quality of life, at least the physical and functional aspects of quality of life in these patients.

To some extent having to go on another drug therapy or having to do something else is a consequence of the treatment failure, and it's there. It's part of that patient's quality of life.

DR. LAIRD: I think you need to separate the

arguments about whether or not to collect the data with what you're going to do with the data. So, I hear what you're saying, that you don't think an intention to treat analysis is appropriate of quality of life data. Whether you thought you were saying that or not, that's the way I interpret what you're saying.

But to me, to make an argument that, gee, I don't think we should collect that data because we can't interpret it, I agree with Diane. You don't know whether or not you can interpret it. Without looking at it, without gathering it, you're leaving yourself vulnerable to all kinds of criticism.

DR. CELLA: Julie?

DR. BEITZ: Yes. What I was going to propose is that there are other settings, for example, in the adjuvant setting --

DR. LAIRD: I'm sorry. I can't hear.

DR. BEITZ: I was going to propose that there are settings, such as the adjuvant setting, where folks get a few months of treatment and then do relatively well for long periods of time, and they could be assessed over time after they've completed the active treatment part.

DR. CELLA: What you're saying there is that the patient is not going to be switching over to another drug or another treatment.

DR. BEITZ: Right. They are survivors, if you will.

DR. CELLA: You have less concern where there's not a switch to another active therapy.

Carol?

DR. LAIRD: You know, I would like to actually raise a related but slightly different protocol issue. I know in a lot of settings I've seen people give the advice that if you're doing, say, a 5-year staggered entry study, so the patients are going to enter for the first 2 years, and then you're going to follow every patient for a minimum of 3 years, that the advice is follow everybody until the end of the 5 years. So, you have from a minimum of 2 years to a maximum of 5 years of follow-up. I can see that if your primary endpoint is time to some event, that that's a desirable way to design your protocol.

But if your outcome of interest is quality of life, so it's a repeated measure, and we're talking about sort of a -- I tend to think of it's maybe a time averaged quantity over the period of interest. The worst thing you can do is have everybody measured for different points of time. And I don't know what the standard cancer protocol looks like.

DR. CELLA: Carol.

DR. MOINPOUR: In Southwest Oncology Group

trials, we do try to follow people. Having defined a follow-up assessment time for quality of life for that particular protocol, in the protocol we say that you're to complete the quality of life assessment schedule for patients through the entire assessment schedule. I do know anecdotally it's more difficult to do that. Once a patient goes off treatment, it's harder to follow him or her.

DR. LAIRD: Yes.

DR. MOINPOUR: I'm intrigued to go back and see if I can actually get any numbers on that.

But that is now a standard part of our protocol, and to me it is just an extension of the intent to treat type analysis because vital status continues to be collected on these patients. It just seems like in terms of comparing treatment arms, you would want to have those patients included for quality of life.

DR. CELLA: Rich?

DR. SCHILSKY: I guess it seems to me that the issue of whether you continue to collect quality of life data after a patient is determined to be a treatment failure depends a lot on the clinical context. For patients who have a metastatic solid tumor and the patient is declared to be a treatment failure, by whatever that definition entails, whether it's progression of disease, symptomatic progression, unacceptable toxicity, whatever,

from the standpoint of evaluating the efficacy of a drug treatment, it seems to me that when the patient is declared to be a treatment failure, you're done. You know that the treatment didn't work.

Now, that may be very different in the adjuvant setting where a patient completes a defined course of therapy and then there may not be another event that occurs ever or maybe not for many, many years later. There it may be relevant to continue to collect quality of life data for some period of time after the treatment is completed, depending again on what your expectations are. Because in the adjuvant setting, you have to cross a line at some point between where you're studying quality of life and you're studying survivorship. That point at which you cross that line depends a lot on what the disease is.

So, at least in my way of thinking about this, you have to really think about the clinical context in which you're being asked to continue the data collection.

DR. LAIRD: Well, I'd like to follow up on that because now I have a question for you. So, what happens when you come to analysis time? You're interested in the course of quality of life over a 3-year period. Somebody was a treatment failure after 1 year. They were declared a treatment failure. You have no more measurements on him. So, it sounds to me like your strategy would be just to

say, well, get that person out of here. They're not relevant. They're a failure, so I don't even care what their quality of life is. Is that what you're saying?

DR. SCHILSKY: I guess I'm thinking of it from a clinical point of view as opposed to an analysis point of view.

DR. LAIRD: Yes, yes. I can see that, but I'm thinking about it from an analysis point of view.

(Laughter.)

DR. SCHILSKY: Sure. Well, that's we have people with different perspectives around the table.

DR. LAIRD: Right.

DR. SCHILSKY: But it does seem to me that once it's clear that the treatment that you've given the patient is not working -- you know it's not working -- the issues of their quality of life subsequent to declaration of treatment failure becomes so complex and so multi-dimensional. Of course, many of these patients with metastatic solid tumors then don't survive very much longer anyway.

DR. LAIRD: Well, right. And I'm not disagreeing with you at all. I think you're probably right. In fact, what you're saying, in fact, would make my job a lot easier because you're sort of saying, once you've identified that the treatment was a failure for this

patient, this patient's quality of life is irrelevant. Then it suggests that when we go to analyze quality of life data, well, we would only include those patients for whom it's relevant.

Of course, the difficulty with that is you're no longer making randomized comparisons because you've removed patients from each treatment group in a non-random fashion. So, it's very much tied to the way in which you decide that patients or treatment therapy in a sort of a main line analysis would need to be first of whether or not what's the proportion of treatment therapies.

DR. FAIRCLOUGH: I'd like to maybe take the other viewpoint, and I would say especially in patients that have a short expectancy, the impact of trying that drug and what happened possibly after they fail and whether their quality of life in their remaining lifetime is better or not -- now, you may not be seeing them as often, but I think it's still very relevant what's happening to that patient as a result of the failure.

And the real impact on quality of life will be often in that failure period rather than the differences that may be associated with very -- I mean, the question is whether you're looking at the quality of life and how it's affected by toxicity of the drug or whether you're looking at an intent to treat that patient with that regimen and

whether that's going to improve their quality of life relative to doing nothing or another regimen.

DR. CELLA: Dr. Chen and then Dr. Temple.

DR. CHEN: My question is from an analysis point of view. Actually my question is how frequently we should assess patients' quality of life. In other words, what's the minimum number of assessments required for, for example, a longitudinal analysis or other type of analysis to obtain a robustness result?

The reason I raise this question is because we reviewed a few NDAs and when a sponsor submits an NDA to us with quality of life data, but when we looked at the data and actually only there were like two or three quality of life measures. So, we had difficulty to like use longitudinal data analysis or other type of analysis. So, then my question is how frequently we should assess the patient's quality of life domain we see in the study period.

DR. CELLA: I think the short answer is it depends.

(Laughter.)

DR. CELLA: And the minimum is two because you need a baseline. And I don't mean to be glib, but it's probably an issue that needs to be drawn out as we get into the detail of what will amount to a set of guidelines or

recommendations in the statistics or analysis section of our task. It really is trial dependent and population dependent.

DR. CHEN: Right. Yes. When we have a meeting with the sponsor, we always have like different assessment schedules. Some of the sponsors submit a protocol with like four or five times assessment, but some of them submit a protocol with only two to three. So, of course, this is a disease dependent question or a treatment dependent question, but from an analysis point of view, if we get too few assessments, actually it's difficult to analyze with a longitudinal method.

But, of course, the other point is if we try to assess data too frequently, then we may increase the missing data. So, we got another problem.

DR. CELLA: Clearly a balance is needed, and we'll be sure to bring some focus in that over time.

Bob, did you want to come back to Diane's comment?

DR. TEMPLE: It seems to me much of the discussion has to do with some lack of clarity about what it is we're trying to find out. If you encounter a patient who has progressed clinically and is on a downhill path, it's not really that interesting to find out the quality of life week to week. I think this is what Rich was saying.

That patient has failed. What you wanted to know is whether you could delay the time at which that disastrous outcome occurs. So, that's interesting. And you also wanted to know what the relation between the toxicity of the treatment perhaps is to the amount of time that you delayed that. But it seems to me that if you're looking at rather small changes over a long period of time, if you confuse that by the disastrous and abrupt downhill turn premorbid, you're just confusing everything.

It really goes to the question I would put if I were asking this at 3 o'clock, what is we're trying to find out here that we don't already know? Since we're measuring survival. We're measuring time to progression. We're measuring symptoms. We're probably measuring performance status. We're measuring all those things. What is the additional thing that we're hoping to get out of all these things? And that's totally interrelated with that question.

DR. CELLA: Let me ask a clinical question to you or to Rich or Stacy or anyone, and that's do we really know enough about a drug that comes in for approval to know that there's not some post, what we call, failure effect. That's one question. So, for example, we say clinically the patients failed. Do we really know that that treatment isn't going to keep having some observable effect on, say,

delaying the further progression of symptoms?

Secondly, if we have asymptomatic patients who progress and continue to be relatively asymptomatic or perhaps some mild symptoms that aren't clinically terribly important, but they're taken off study because they've radiographically progressed, 25, 50 percent enlargement of measured tumor, might there be some reason to believe that those patients might still get some symptom benefit after the time they're taken off the trial?

So, these are two examples of times where even the clinician would, I think, have some interest in knowing might there be some benefit in looking at this person's symptom status, quality of life, functional abilities after we've taken them off a treatment where, by the response data, they have to come off?

DR. TEMPLE: Rick is going to give you a better answer on some of that.

But usually when someone progresses radiographically or in some other way the treatment is stopped. Now, could it have some residual effect later? I think it's really hard to answer that question. There might be therapies that someone would, for some theoretical reason, continue even after there had been some tumor growth. Then it would be sensible to ask those questions.

But on the specific question of the patient is

failing, going rapidly down hill, it seems to me it confuses the issue to get a lot of data at that point because their quality of life is definitely awful. Is that really what you wanted to know? Did you want to know that radiographic and symptomatic progression impairs quality of life? Is that something we didn't know?

DR. CELLA: Okay. Rick and then Kay.

DR. PAZDUR: One small point. With some of the newer drugs that are coming out, the cytostatic drugs, this may have a definite impact where you aren't going to see the effect of a drug for some time. Obviously you may get tumor growth before you have a chance for the drug to actually work. That's being actually written in some of these studies. So, I think perhaps in that clinical setting in drugs that don't have a classical action of tumor reduction or shrinkage of tumor where we're trying to look at more of a stabilization of disease, that concept of this discrepancy between radiographic progression and benefit to the patient may be in a desperate situation.

DR. CELLA: Kay and then Claire.

DR. DICKERSIN: Bob, I can understand that to a clinician the quality of life outcomes might not be clinically interesting, but I certainly could see how the drug would have effects that would compete with the bad effects that you're going down hill in terms of dying that

would be important to the patient. If it were just about clinical outcomes, then we wouldn't be talking about quality of life. I just don't understand that once the treatment isn't working, you aren't interested in quality of life data.

DR. TEMPLE: It's not that you're not interested in it. It's that you already know.

DR. DICKERSIN: No. You know clinically. You don't know how the patient is feeling.

DR. TEMPLE: You have a cytotoxic agent whose only effect that you know about, apart from adverse effects, is to kill tumor cells or slow their growth or something like that. You're now in a position where the tumor is growing again rapidly enough so that it has met whatever standard you have for progression. I guess it's not inconceivable that the drug is also an antidepressant, but you're not going to keep giving it and it's clear what, apart from the fact that things are deteriorating rapidly, you're going to see. That's what you'll find: The quality of life is getting worse because all of the things you're worried about and we're trying to forestall are now beginning. So, what is it that you hope to find?

DR. DICKERSIN: Well, quality of life isn't a single global measure. There are all different aspects, some of it having to do with the disease itself and the

disease progression and some having to do with the effects of the drug.

DR. TEMPLE: Well, a drug can make you feel worse. It's unusual for a drug that's used to treat tumors to make you feel better, if it's an anticancer drug. If it's an antidepressant, that's a different question. But it was designed to shrink the tumor and make it not grow. What's your hypothesis about what the benefit is going to be apart from doing something to the tumor and the results of that effect on the tumor? What are you imagining even that you'd be looking for?

DR. CELLA: Well, is there a quick answer? Because we have a list of four that are waiting.

DR. TEMPLE: You could definitely have more toxicity, but this is sort of over then because once you progress, people take you off the drug. You could have residual toxicity.

DR. DICKERSIN: What if you've had some of your functions disturbed so that now you have diarrhea or fecal incontinence, et cetera, et cetera?

DR. FAIRCLOUGH: Or permanent cardiovascular damage.

DR. TEMPLE: Yes. You can be worse off. But you would have had that already. You're going to see that. By now people have stopped the drug. You're interested in

late developing toxicity.

DR. CELLA: I guess just a short answer I'd put and then I'll turn to Claire would be, as I alluded to earlier, some of the things you would or could see that you otherwise wouldn't know and would assume have to do with the rate of the progression and the severity of the progression of symptoms that may differ by treatment arm as a function of the treatment that you're looking at. You're evaluating a treatment and you want to know if it helps people. And to the extent that there's some residual effect that occurs after you've taken that patient off study, there's an argument for gathering more patient-reported information so that you can determine if that rate of change differs.

It's complicated obviously by the fact that they go on to other treatments, and if they're likely to go on to different treatments as a function of treatment arm so that that's bound into the randomization, then you have a potential problem of interpretation there.

But I think that Dr. Laird's point was that that's a decision she's recommending making later. That's not a decision of collecting the data. It's a decision about how to handle it and analyze it. I agree that those are things that are worth keeping separate.

Claire.

DR. GNECCO: Thank you. I think we lost Professor Laird, didn't we?

DR. CELLA: Yes.

DR. GNECCO: Perhaps Diane can help me out.

I was thinking, getting back to this problem with data collection when you do have treatment failures, if it's truly impossible to collect the data, what about these newer, very sophisticated, multiple imputation techniques that Professor Laird alluded to? And when I'm talking about our model based prediction and propensity scores, there's a software package, which should remain nameless I think in this setting, that does implement these things, and it does allow you to make use of the ancillary information, the ancillary covariates that Diane mentioned in her talk. Might there be a place for that approach in the analysis? Diane, would you like to take that on?

DR. FAIRCLOUGH: Yes, partially. One of the things that Dr. Laird said very clearly but it might have slipped by very quickly is that it's only when you have some measurement on some of the patients and you can assume that the other patients that have missing data that may be experiencing a certain grade of toxicity, that it's impacting those groups equally.

I think, if we're talking about the same package actually, that package makes very weak assumptions

about the missing data. It's almost missing completely at random given the strata you're assigned to by the propensity scores. I worry that it's too easy to use without a lot of good clinical background. If you don't have those other indicators and they don't get into the model, you're just obscuring the problem. But it's definitely a tool.

I use some explicit model based methods, and they definitely help inform me about what may be happening in that data.

But if by protocol design you stop the assessment, conditional on a certain event, then you have absolutely no information. You can only guess, and there's nothing to help you do the imputation. You're doing the imputation in the absence of information.

DR. GNECCO: That's been our impression too, but we are, at least in the Center for Biologics, very recently seeing more sponsors proposing using this package. I do think it has its use but with all the caveats that you gave us. Thank you so much.

DR. CELLA: Jody?

DR. PELUSI: When I look at this issue, I come from a clinical perspective and trying to look at what is the charge of the FDA and then how do I translate it out in the real world. To me the charge of the ODAC is to look at

safety and to basically say is this drug safe to use and what do we as clinicians need to know to provide that information and support to make it worthwhile in someone's life.

But when we see the disease failures on a clinical trial, it is good to keep the quality of life information going. But many of our patients aren't going to be there, and so we're not going to have that long-term follow-up.

My concern is many of these things are going to go on the market because we're not going to know the impact of quality of life right in that study and it's going to take us years. A lot of those patients won't be there to collect that data later.

Is there -- and I know the statisticians will probably have chest pain at this point -- but when you look at we put it into practice and then we have another whole cohort of individuals that are on this treatment that actually may be able to supply us with that information because the original ones may not be with us.

An example I would give is, as we sit in small community hospitals and we say, okay, now we have this regime, but what we're starting to see is this toxicity which is significantly impacting the quality of life where a patient says, if I would have known this, I don't know if

I would have gone on that treatment. Well, those individuals originally may not be there.

So, the question becomes, do we have to gather some of that data from another whole group of patients that are not necessarily linked to the original study? Do we ask the clinicians to start to look at trends in terms of side effects that can ultimately look at quality of life as well? I'm just afraid that with the follow-up issues we're going to miss a lot of that data, and that's significant in terms of long term.

DR. CELLA: Thank you.

Rich?

DR. SCHILSKY: At the risk of prolonging the debate, I guess I want to try to clarify a few issues.

First of all, I think as physicians caring for patients, we're always interested in the quality of the patient's life for as long as they live. So, we shouldn't be misunderstood and have anyone go away from this meeting thinking that we're not interested in the patient's quality of life after they progress on a therapy.

That's not what we're talking about here. We're talking about how informative is collecting additional quality of life data in order to make a regulatory decision. Again, it seems to me that that has to be put in the appropriate clinical context.

For patients with metastatic solid tumors, the average survival after failure on initial chemotherapy is in the realm of 3 to 6 months. So, for most of those patients, there's not going to be much opportunity to collect much additional information.

Secondly, it's not clear that collecting that information is actually going to be informative with respect to making a regulatory decision on whether the drug should be approved or not in that kind of an indication.

Third, of course, there are well-established mechanisms existing in the country for collecting adverse event data even on marketed drugs so that if the drug is out there and there are adverse events occurring, they're likely to be reported.

Fourth, it seems to me where the quality of life information long-term is probably the most important is in the setting where you expect long-term survivors, and that's where there's also the greatest opportunity to obtain that information in a reliable fashion. When a drug is effective in the metastatic disease setting, frequently it's then used in the adjuvant disease setting, in which case the therapy is typically given for a defined course of time. The patient is then removed from the therapy. The patient then has a relatively long life expectancy. There's plenty of opportunity to observe the patient and