Initiative on Methods, Measurement, and Pain Assessment in Clinical Trials

IMMPACT-XVII

Assessment of Physical Function in Analgesic Clinical Trials

Thursday, April 17, 2014
Westin Georgetown
Washington, DC

Carole Ludwig
Transcription Services
141 East Third Street #3E
New York, New York 10009
Phone: (212) 420-0771
Fax: (212) 420-6007

TABLE OF CONTENTS

ACTTION and OMERACT Welcome and Introductions

DENNIS TURK: Good morning, welcome to all of you who are drifting in. My name is Dennis Turk, I'm the co-chair of the IMMPACT group which you will hear more about shortly, and I want to welcome you on behalf of Bob Dworkin and myself, and Phil Mease from OMERACT to attend the first combined meeting or joint meeting, if you will, of IMMPACT and OMERACT, and it's a tremendous delight for us to have you all here with us to address what we think would be a very interesting program, very interesting questions that concerns to all of us who are involved with clinical research, clinical trials, as well as clinical practice for those of you in clinical practice. So we really are delighted.

I want to especially thank those of you that came long distances, especially those who came from Europe, thank you so much, we understand about jetlag. I come from Seattle, I have jetlag sort of in the opposite direction, so you'll understand if I'm not totally functional.

A couple of housekeeping details that we want to cover. Just so you know this, there is no internet access in this particular room so if you need to use internet access you probably have to leave the room. Please mute your cell phones, put them on something, but don't have

them going off. And it's interesting because every meeting I've ever gone to that statement is made and I've never been to a meeting yet where somebody's cell phone didn't go off. So let's see if by any chance we can pass that today.

The microphones, you'll notice that they are, number, one, acutely sensitive, so if you touch your desktop or say anything the light will come on, you don't have to push any buttons. If you want to speak, the light will come on when you start to speak it will go off automatically. But be careful because if you whisper to your neighbor we're all going to know what you're whispering. So unless you want us to know about that, don't do it, okay?

I'll let you know that the restrooms are to my left, your right, down the hallway, past where you came off the elevators, if you came off the main elevators. The lunch will take place upstairs, the coffee break is going to be taken outside here. Dinner will also be in this hotel and you'll get information about where that is, it's also upstairs.

Registration, the checkout time tomorrow, is not till twelve o'clock tomorrow noon, so at roughly either the coffee break tomorrow if you don't want to do it sooner, is probably when you ought to think about checking out of your

room, if you have a room, for those that have rooms.

Let's see, I want to thank a couple of people that most of you have either met physically or at least had email correspondence with, they are not in the room right this minute, but that's Valorie Thompson who just opened the door right there, and Andrea Speckin, they've been the organizers. They are here to help you in any way that they can, so if you have questions about transportation, hotels, beautiful things to do in Washington, DC, in your off times, they can help you with that. And if they are in here I will thank them profusely because this meeting really can't happen without them helping us and cooperating.

And as I already thanked my colleagues from OMERACT as well as IMMPACT steering committee who have been helpful to us in organizing all of the meetings that we've had, but especially this meeting, I thank you and I hope that for our OMERACT friends this may be an opportunity that in future meetings that we might find there are mutual benefits to doing them together. I know that you'll hear from Phil Mease who will tell you a little bit about OMERACT.

For those that don't know about IMMPACT I'm going to just throw up a couple of slides to give you a bit of background and then Dr. Mease, Phil Mease, will do the same

thing. So this is what you are here for, the IMMPACT meeting. IMMPACT stands for the Initiative on Methods, Measurement, and Pain Assessment in Clinical Trials. OMERACT, which is Outcome Measures in Rheumatology, should be probably clinical trials, the CT for OMERACT. This is the 17th IMMPACT meeting, the first joint meeting that I mentioned. It's going to be specifically focusing on assessment of physical functioning in clinical trials, and this has been a topic that at least for the IMMPACT group has been a concern for a long time. We focus a lot on patient-reported outcome measures, self-report measures, even some biomarkers, but really when we talk about outcomes we recommended in the early IMMPACT papers physical function should be considered in a clinical trial.

We haven't really spent a whole lot of time talking about how do you evaluate physical functioning. And that can go everywhere from self report measures to performance measures, to observational measures, there's a lot of ways that you can do that, but we have really not given much attention. So from the IMMPACT side this will be a very useful meeting for us to have to cover that topic.

What is IMMPACT - as I already mentioned, it's an international consortium of participants from academic research, governmental agencies, both from the United

States and from Europe, industry consulting organizations, and consumer advocates. I'm using the word consumer advocate or patient advocate, or patient support. I am purposely using the word consumer advocate and not patient support because of one of our colleagues sitting in the back who keeps reminding me that people who have pain problems shouldn't be talked about as patients, they should be talked about as people. So if I slip, Penney, I know you'll catch me as you always do.

We are very much interested in also something called ACTTION, which I'll tell you a little bit about towards the end of this presentation because IMMPACT has now been merged within ACTTION. The mission of IMMPACT initially was to develop consensus recommendations for improving the design, execution and interpretation of clinical trials in the treatment of pain, fairly specific and fairly clear. We're focusing our efforts on research methods, methodologies, ways of improving clinical trials, what the ultimate endpoint hopefully being that if we in fact can improve those trials, improve the way that we do research, we can actually expedite the process of getting better information which hopefully can be useful as we start thinking about better treatments for patients, or people who have pain problems, they're patients when we

treat them, Penney.

Our background, who is IMMPACT? To date, including this meeting, we've had 261 different participants at 17 different meetings, a number of you have been, at least the IMMPACT people have been to more than one of those meetings. We've had 146 academic and clinical scientists who attended these meetings. They've come from the United States, from Europe, you see the countries, from Canada, from Australia, so we have not as yet had anybody from Asia, so that's where you from OMERACT are ahead of us because you've involved people from Asia as well.

We have had representatives coming from 82 different academic and clinical institutions worldwide. There have been 71 representatives coming from 38 different pharmaceutical companies who have provided support to IMMPACT when it was existing independent of ACTTION. There are four representatives from consumer advocacy/support groups, patient support groups, take your choice of what term you'd like to use for them. And we have had over time 36 participants from different governmental agencies. The agencies, I just listed them for you so they come from a range of different NIH institutes, they come from Department of Defense, from the DEA, as observers. We've had people come from SAMSA, from the FDA, from, as I said,

NIH, from the VA, et cetera. So you can sort of see, and there have been a lot of representatives who have come, who have participated to the extent that they're able to, have contributed tremendously, and we hope that the meetings have been useful and productive for them to learn about the kinds of issues that we're concerned about, we discuss.

These are the different IMMPACT meetings. This is the first, there will be two slides, these are the different meetings, they go everywhere from outcome domains, outcome measures, patient-reported outcomes, patient improvements, pediatric trials, confirmatory drug trials, acute pain trials, proof of concept, et cetera. So it's a whole range of different topics. Each year, roughly once a year, but there have been a couple of years where there were two. Roughly we cover one of these particular topics, there's a steering committee who at the end of this meeting, at the end of every one of our meetings we talk about what are the next topics that we think would be worthwhile covering, we come up with a set of those topics, we discuss it with the steering committee for IMMPACT and then we pick a topic that we think would be most appropriate, then the steering committee helps us as we move forward in identifying speakers, background papers, and dates that these meetings can occur.

So those are some of the meetings, there are two others that were not on that slide, the most recent ones were on biomarkers and on phenotyping patients in clinical trials. And this particular meeting that you're all attending is the assessment of physical function in analgesic clinical trials. So that's what we're here for.

These are the people who are at this meeting. I think it's the most recent list I had. I highlighted in yellow those who are going to be speaking. So one of the things we have done, in the past IMMPACT meetings have run anywhere from the size of let's say 25 to 45, this room is getting pretty big so to go around and have everybody say who they are and where they're from takes a lot of time. So what I would like you to do in lieu of that is the first time you have a question, the first time you speak, please introduce yourself to us as well as where you are coming from, what facility you're working at, or anything you want to say, but rather than trying to go around the room which would just take too much time.

You can see we have people from different countries, we have people who are coming from all over the United States as well as from Canada, as well as from Europe, and we're delighted to have all of you here. And again, I thank all of you that come from a long

distance.

This is the website for IMMPACT. If any of you are not familiar with IMMPACT and want to learn more about past meetings, on this website you'll notice that every one of the past meetings has listed all the people who were available attending those meetings, any background papers that were presented, slide presentations, it's all available, it's hopefully transparent for you to see. And who sponsored the different meetings, that's all available to anybody who chooses to want to look at that

Now I mentioned in an earlier slide that IMMPACT which started in 2002/2003, has since that time been merged, subsumed within ACTTION. ACTTION stands for Analgesic, Anesthetic and Addiction Clinical Trial Translations, Innovations, Opportunities, & Networks. So that's what ACTTION stands for. You'll notice we like acronyms, you'll see lots of these. One of our favorite things to do is sit around coming up with acronyms.

Now the mission of ACTTION, which is a public-private partnership with the United States Food and Drug Administration, is to identify, prioritize, sponsor, coordinate and promote innovative activities, with a special interest in optimizing clinical trials that will help expedite the discovery and development of improved

analgesic, anesthetic and addiction treatments for the benefit of the public health. And I have to say Dr. Bob Rappaport, I saw him in the back of the room there, Dr. Rappaport was the motivator for why there was a need and a value in having a public-private partnership. He strongly encouraged and worked with the FDA to move forward on this initiative. We were fortunate to have IMMPACT imbedded within the ACTTION initiative. ACTTION is much broader than IMMPACT in that in addition to having these meetings, it also sponsors a whole range of other projects. You can go to the ACTTION website which is at the bottom which tells you the different working groups. I think the last time it was 9 or 10 different working groups, Bob, that are looking at different aspects of improving clinical trials, educating consumers about, one of the most recent ones that Penney Cowan from the American Chronic Pain Associations is working with us to try to educate the public about what are clinical trials and to educate them on why they might be interested and why they should consider becoming involved in clinical trials. We also are working with Penney in trying to find ways to help the public understand what the drug development process is from molecules to actually marketing so that people understand what goes into these things.

Other projects are involved with trying to find ways to educate participants in studies or patients in clinical trials, but how to even fill out questionnaires. We've done some focus groups and interestingly found out, much to our surprise, that when we ask patients, people to give us their responses to what their average pain was, they didn't understand what we meant by average. So almost every questionnaire that we looked at asked about the average pain and they said well do you mean average from when I woke up this morning, what about when I was sleeping, should I average in the zeros from when I was sleeping, what do I do with that.

So those kinds of focus groups which the OMERACT people are very good at involving consumers, patients, if you will, in their projects, ACTTION has learned a lot more and IMMPACT is learning from ACTTION about the importance of making sure that we understand what the consumer and user, the patient, and it could be the provider, how they understand what we're doing in these particular trials. So those are the kinds of working groups looking at secondary analysis of published data to improve the outcomes, to improve the, it's so much the outcome, to improve the design of the design of the studies. We're not so much interested in looking at can we get a better outcome, but

rather can we do a better job of designing the trials so that when we see outcomes we understand better what do they mean and how to interpret that information. Can we use existing datasets that are published and then can we use those datasets to learn about the design of clinical studies and to improve the design of the clinical studies, with the cooperation of the FDA when possible. We also work to try to get access to some raw data from the FDA's databases with the approval of the appropriate companies that have provided those data to try to learn more about not the outcome, but can we learn something about how to design studies that will speed up process and give us better outcome, better understanding of the process.

So that's what we've been all about. Now I want to turn this over to Phil Mease who is from OMERACT who is going to, I assume, in addition to welcome you, give you a little bit of background for those from the IMMPACT side who may not know as much about OMERACT, to explain pretty much what OMERACT Parallel is and what their interests are in working with us. So, Phil, thanks so much for cooperating with us.

PHILIP MEASE: So in the next few minutes my role is to introduce you to OMERACT, for those that don't know about its history and function, and then specifically about

the chronic pain working group within OMERACT which is the group from within OMERACT that is meeting with IMMPACT here.

OMERACT has been in existence since 1992. It holds meetings every other year in a variety of places around the globe partly related to its charter with WHO. And it is closely connected, as well, with the Cochrane initiative. The attendance at a typical OMERACT meeting is about 250 people but there are literally thousands of OMERACTers that are spread about the world now and who have been involved in one way or another and there are hundreds of publications from the proceedings, as well as the output of the working groups. There is an executive and steering committee and then decentralized working groups. There are representatives from academia, clinical investigators, industry, regulatory agencies, that are involved.

There are several key outcomes that come from OMERACT initiatives. One is studying a disease in rheumatology and trying to define what is the core set of that disease that defines the disease for a patient, from a physical/biological point of view, from the impact of the disease on the patient, to the impact on society, et cetera. And so there is an exercise that OMERACT goes through to define the key elements that define a disease.

Once that's done, then one can look at the various outcome measures that are used to assess those elements, and determine whether or not those outcome measures are valid, represent the truth, discriminate, for example, in a clinical trial between placebo and a treatment group, and are feasible to do. And then another activity that the working groups do is to develop then responder analyses, definition of state, response, minimal clinical important disease, remission, low disease activity and so forth. And then the research groups which are doing their work between the meetings have an ongoing research agenda for their activities.

I've listed in the first sub bullet some of the disease states that are represented underneath the OMERACT structure, ranging from rheumatoid arthritis to osteoarthritis to vasculitis, to the various spondyloarthritis conditions, to connective tissue disease, interstitial lung disease and so on. There are also specific working groups that are dedicated to imaging, be it x-ray or MRI or ultrasound, and the various issues that come up in standardizing imaging outcomes. And then there are what I've called special issues here which include addressing the patient perspective, fatigue, participation in important life activities, interaction with the

International Classification of Functioning, as well as PROMIS.

And I want to emphasize the role of the patients within OMERACT. This has been an increasing representation and patients really work with OMERACT in two different ways. One is the identification of what we call patient researchers. These are people that actively participate with the working groups just like any other member of the working group. The person may be a patient with a disease but also say they have a background as a research psychologist or have some special interest in pursuing methodological research. And so those kinds of individuals will often take part in the research, development, analysis and writing of the various activities of the group.

We also use patients, or consumers as I'm learning from Dennis and Penney, that we use them in the form of focus groups or work with them in the form of focus groups. So, for example, when defining the concept of flare in rheumatoid arthritis, one might have focus groups of 20 individuals with rheumatoid arthritis and listen to them describe what they think of as the concept of flare of the disease, write it down and then go through exercises like Delphi exercises to prioritize what are the key elements that define flare. And so increasingly in both ways

patients work with OMERACT investigators.

Also there's an emphasis on engagement of fellows. These are either actual fellows in rheumatology or junior faculty. And these are encouraged to participate as rapporteurs within the various meetings that we have and then are engaged in writing the proceedings.

This is a slide that I borrowed from Vibeke Strand that gives an example of the relationships of OMERACT with other organizations. So, for example, in the conduct of our work in rheumatoid arthritis we work closely with EULAR, the European Rheumatology Association, as well as ACR, the American College of Rheumatology. And as you can see, in other clinical areas there's a close relationship with the research organizations such as OARSI in osteoarthritis. This meeting is an example of a meeting with IMMPACT in the area of pain. We also work closely with the ICF and PROMIS.

On the left-hand side of the slide I've mentioned some "baby" OMERACTs that are developing. One of the ones that I happen to be working with very closely is one called IDIOM, International Dermatology Outcome Measures, and this is a group of people interested in the realm of psoriasis who would like to take the same methodological rigor that is used in OMERACT to the world of dermatology. So they're

starting with psoriasis, looking at what defines the clinical domains of psoriasis, what is the performance characteristics of the outcome measures used in clinical trials, development of a responder in disease and so on. And other disciplines are doing the same.

Here is the so-called OMERACT filter. So whether you're looking at constructing the core set, the clinical core set, or looking at responder analysis development, at each level you have to ask the question is this true, does it fulfill face, content, construct and criterion validity. Does it discriminate? Is an outcome measure, for example, that's developed, reliable, reproducible and sensitive to change? And can it be feasibly done in a clinical trial? So this is a filter by which we always are looking at the various measures that are developed.

In the last two OMERACT meetings, there has been a sort of refocus on how one goes about developing both core sets as well as development of outcome measures. This is called OMERACT 2.0. I won't get into this detail but the point here is that it broadens the way in which one looks at the impact of disease on the individual, their family and society. So not only looking at the right hand side of the slide which is the traditional way of addressing a disease and its impact on the body, and the way one

measures the disease state and change with treatment, but also looking at the impact that the disease has on a patient's life and the resource use within the society, as well as taking into account adverse events that may occur during the course of a trial. And particular context that a trial is done in, for example, is it representative of gender, age, ethnicity and so forth that is going to be applicable and generalizable. And so here is a roadmap that has been developed for developing a particular domain set.

So this is my last slide and so just to say that the chronic pain working group has been recently initiated within OMERACT but has a long history starting with the fibromyalgia working group. And there are several people here such as Dan Clauw, Dave Williams, who had been working with IMMPACT for a long time that were part of this working group. And the key originators for the chronic pain working group are noted here, Lee Simon, myself, Peter Tugwell who is here, Vibeke Strand, Bob Dworkin and Dennis, and the participants include members from multiple different OMERACT working groups. So people that are particularly interested in pain, for example, in osteoarthritis, for example, Philip Conaghan who is here at this meeting, or rheumatoid arthritis, fibromyalgia, or spondyloarthritis and so on.

We have two fellows here, Kristine Phillips and Ann Taylor, maybe you all could stand up just a moment because you'll be -- thanks, very much. So Kristine is a junior faculty at the University of Michigan in Dan Clauw's group, Ann Taylor works with Ernest Choy in Cardiff, and are going to be presenting and also rapporteurs in some of our discussions and being scribes for what we're doing here. And administrative support is supplied by Valorie Thompson.

So the purpose of this working group is to define a core set of pain, the pain domain in various rheumatic diseases, to evaluate and standardize measures across rheumatologic conditions, and one of the issues that we deal with quite a bit as a challenge or problem within assessment of rheumatic diseases, is the whole issue of defining what in historical terms would have been called peripheral pain versus central pain, if you will. For example, in a rheumatoid arthritis population, approximately 20 percent of patients will have what we call central sensitization or central pain or you might use the F word and call it secondary fibromyalgia. And this ends up confounding our analysis of the pure impact of rheumatoid arthritis, it's part of the disease spectrum but it influences all of our measures of outcomes in rheumatoid

arthritis.

So going about being able to assess and identify the contribution of central pain I think is an important area, it's just one example of many. And then the issue that we're tackling in this meeting which is the whole impact of pain on physical function.

So we're having this meeting today and tomorrow, in Budapest next month we will be having a pain pre-meeting at the OMERACT meeting, and then a special interest group meeting within the context of the Budapest meeting.

So with that I'll stop and turn it back over to Dennis. Any questions about any of this? Yes.

JOHN FARRAR: Quick question. So the central versus peripheral was the basis of or is the basis of a number of meetings that IMMPACT has had, and clearly is a huge issue in trying to phenotype patients with regards to that. So I wanted to ask a question relative to your definition of chronic pain. One of the decisions that I think has been generally accepted is the concept that at least in some circumstances, chronic pain is considered a disease unto itself. I'm a neurologist, I'll admit that up front, and I have a bias, pain is basically neurologic from the get go. But I just wondered if that was sort of the approach that this pain group was taking, and then I

actually wanted to ask a second very short question, Dennis is shaking his head, I'll talk to you later. But I just wondered if you could help me with understanding.

MEASE: So part of the reason I used the word historical concept was just the point that you're making, that we increasingly appreciate that pain is a disease state unto itself and there are peripheral and central, if you will, contributions to that, and it's very hard to separate. I saw Lee grinning there because there's a lot of debate and controversy in all of this.

One of the things that we are aware of is that when you turn to the rheumatology community, they are ages behind the way this group thinks and really have a hard time accepting that there may be genetic and central mechanisms that are at play that they need to think about and address. And so part of our role is to, if you will, take the neurophysiological understanding that's been growing amongst a group like this and bring that to rheumatologists so that they understand.

TURK: Thank you, John, I'm sorry I'm stifling you, only because I want to save those discussions. And by the way, you broke the rule, that's Dr. John Farrar from the University of Pennsylvania who is a card carrying neurologist and current and former IMMPACTer.

One thing that I forgot to mention is that in addition to having meetings, what IMMPACT has also been very much involved with as is OMERACT, are making sure that we disseminate information, that we develop publications based on our considerations so that it's not just the people in the room that benefit, but that it actually benefits the field. So we attempt to publish our manuscript papers in the most appropriate kinds of journals we can. As I showed you the IMMPACT website, if you want to see any of those particular publications, we try to have at least one publication per meeting.

That means this meeting, one of the expectations is, the reason we have scribes, the reason that we have people, the way that we schedule the program, the way that we did, was so that at the end of the day we will be able to develop some considerations, some discussion points, some research agenda that will appear in a publication that we will submit to whatever we consider is an appropriate journal. All of you will be invited to be authors of that manuscript, at least assuming there is only one but there could be more. You will be asked to be authors.

What happens, so you'll understand the process because you're going to say how do you get a manuscript with 40+ people? What we do is our scribes, our

rapporteurs draft something up. We then circulate it to the steering committee to refine, to clean up. We then will send it out to each of you to review, to give us comments on, please, please, please, when we send it to you, try to give us a reasonable turnaround because we don't want it to take five years for manuscripts to appear. And I'm going to say please yet again, that we'll take your comments, we'll try to imbed them as best we can, depending upon how extensive they are you might see another round of this for your consideration. Anyone who chooses not to participate can do so, there is no requirement, but we strongly encourage all of you who participate in the meeting to participate on those manuscripts.

The IMMPACT papers, and I can't speak for OMERACT, have had quite a scientific impact. They have been cited over 1,500 times since the first publication. There have been approximately 17 or 18, maybe 20 publications have come out. The most gratifying thing to me is that they've been cited in over 500 different scientific journals going everywhere from additional medicine to veterinary medicine, women's health. Ask me why veterinary medicine, I haven't got a clue, but we could ask, John could tell me about that since he's been involved with some of those. But the idea is that we want to have a scientific

effect. We don't want it just to be we talk about it among ourselves, but we want to disseminate that information.

Assuming that we come up with something useful and valuable, we want to make sure that those considerations get put out there so that people who are doing clinical trials, this is purposely the idea is for studies, doing research, so if you are going back to your particular facility n Monday morning, hopefully you'll learn something at this meeting which may be useful to you as you think about designing your clinical trial.

So all of you will be actively involved, we've built this schedule, the program, such that there's a lot of time for discussion. You'll notice that when you look at the schedule which is in front of you. You'll also notice that the schedule is a little flexible because timing isn't exactly to the minute, we understand that and there may be some last minute shifting and changing things around.

So speak here. When you first speak the first time, please introduce yourselves because not everybody will know everybody else. So as John didn't do for us. There will also be coffee breaks, lunch periods, plenty of time for discussion and we hope that the discussions actually are as important, if not even more important, than the actual presentations. The presentation speakers that

you're going to hear are to stimulate discussion, to stimulate consideration, to have debates, to have different perspectives. We have neurologists, we have anesthesiologists, we have psychologists, we have health scientists of all different types who are in the room, and they come with a different perspective and that's helpful for us to be thinking about that.

So that's where we're going. Before I actually start introducing our first speaker, any questions about what the purpose of this meeting is, what we hope the endpoint is going to be, and how we evolved the program the way we did? I'll let Bob take any difficult questions, the easy questions I'll take.

Crystal clear to everybody, you know your responsibilities and you -- Bob? Who are you?

ROBERT DWORKIN: I'm Bob Dworkin from University of Rochester. One thing I think we neglected to say is we are taping this meeting from beginning to end and a transcript will be prepared from the tape, and that transcript will be made available on the IMMPACT website. So if you are opposed to that approach to doing things, you should leave now because we're not going to try and come up with a consensus about whether that's what we're doing, we're doing it, and this really, not to be too facetious,

this is in the interest of as great a transparency as possible.

We don't know whether anyone is going to go through the trouble of reading what will probably be a 200 page transcript, but it will be available on the IMMPACT website in a couple of months.

MALE VOICE: Will you redact the side comments?

TURK: I told you, these are very sensitive microphones, so if you burp it's going to be in there, so pay attention.

DWORKIN: Someone will read the transcript before it is posted.

TURK: The other option is if you don't want to have anything you say on the transcript, don't speak. That's obviously not something that we want to have you do. We do want to, as I said, we want to disseminate as widely the information as we can. It is not a secret, we want the information out there to help people. That's what we're all about.

Other questions besides Bob Dworkin from the University of Rochester? Good, that's how you introduce yourself. Okay, then let's start with the formal meeting and each speaker will decide for him or herself whether they want questions during the speaking. Usually I think

it's best not to ask questions in the middle when people are speaking because it distracts them, but there is plenty of time for questions for discussion. So let me start the program with our first speaker. We're delighted to have Dr. Ashley Slagle from the Food and Drug Administration. She's going to be talking about the FDA's considerations in development of and the selection of outcome measures. And you'll see throughout this presentation, throughout the day, we also have the perspectives from different agencies, as well as from academics, because we want to have people speaking with each other. So Ashley, I'll give you the group.

FDA Recommendations for the Development or Selection of Outcome Measures

ASHLEY SLAGLE: Good morning. So to avoid getting in trouble, let me announce myself, I'm Ashley Slagle from the Study Endpoints and Labeling Development Staff in the Office of New Drugs at the FDA, and my role is to give you a regulatory perspective on the approach to outcome measure development or selection. So of course my disclaimer slide, the views here are my own and do not necessarily represent the official position of the FDA.

So before we talk about the details of measure development or selection, I want to step back and think about the broader context of what we're trying to do in clinical trials. So ultimately we seek to evaluate treatment benefit so that the drug has some positive impact on something that is important to patients or consumers in their daily lives. So how long they live, how they feel, or how they function in daily life.

So we use outcome assessments in clinical trials, the purpose of an outcome assessment is to determine whether or not a drug provides treatment benefit. So based on that thing, or the concept of interest that was measured by our outcome assessment in a clinical trial, we can reach a conclusion of treatment benefit. And then we must

describe that benefit in labeling in a way that is not false or misleading.

So there are several types or outcome assessments we can use to assess treatment benefit, of course survival, biomarker and clinical outcome assessments. So survival is pretty straightforward. In the case where survival is not readily measured in clinical trials, or we want to understand additional supporting information, we might turn to biomarkers or clinical outcome assessments.

And clinical outcome assessments are any outcome assessment that depends on human judgment, motivation or participation, and these will be the focus of most of my discussion. And these can include patient reported outcome assessments, clinician reported outcome assessments, observer reported outcome assessments, and performance outcome measures.

So how do we determine which type of clinical outcome assessment is best for a particular clinical trial or context of use. Well we're always interested in the patient or consumer perspective, but remember that in our era of patient focused drug development that doesn't always necessarily mean that we need to use a patient reported outcome assessment as an endpoint in clinical trials to evaluate treatment benefit. It means that we need to think

about what is important from the patient perspective in a particular population, and then determine how best to measure that so that we can understand treatment benefit in a particular clinical trial.

So it may be that we're interested in symptoms or physical function in a population that can report for themselves, in this case a patient reported outcome assessment may be the best option. If clinical judgment is needed to interpret an observation, then a clinician reported outcome assessment could be chosen. If an observable behavior in daily life is being assessed in a population that can't report for themselves, then observer reported outcome may be appropriate. And in some cases if we want to observe an actual demonstration of some defined task in the clinic of functional performance, we would use a performance outcome measure.

So let's talk a little bit more about treatment benefits. We think about treatment benefit in term of direct and indirect evidence. Direct evidence of treatment benefit is derived from studies with endpoints that measure survival or how patients feel or function in daily life. Indirect evidence of treatment benefit is derived from studies with endpoints that measure other things that are related to how patients survive, feel or function.

So it might be helpful to think about this in terms of a continuum of direct and indirect evidence of treatment benefit. Depending on how indirect something is, the more evidence we may need. So we consider performance measures like the 6 minute walk test to be somewhat indirect because they are not measuring how people feel or function in their daily life, but are intended to closely approximate how patients feel or function in daily life.

With these types of performance measures, it is critical to understand what the performance test is actually measuring, what the score represents, and what all that means in terms of meaningful treatment benefit to patients in their daily lives.

Biomarkers are surrogates for treatment benefit and are at the far indirect side or the far right side of this continuum. Therefore, we need very strong evidence showing that the biomarker predicts or relates to clinical treatment benefits and how patients feel, function or survive.

So surrogates are biomarkers within an existing, well established link of clinical benefit, can support endpoints in traditional approval. So, for example, blood pressure. Biomarkers are surrogates without that existing evidence to their link to meaningful treatment benefit, but

are reasonably likely to predict clinical benefit, might be able to support an approval through the accelerated approval pathway, with the requirement that post approval studies are completed to then confirm the link between the biomarker and the expected clinical benefit.

So also when we think about evaluating treatment benefit, we recommend that the core disease defining concepts be assessed first, and considered before we think about more downstream effects. So, for example, thinking about the assessment of physical function in patients with pain, we might think about knee pain and specific knee pain related physical functioning, such as the ability to bend the knee which is related to walking, as primary, secondary endpoints in clinical trials.

As we move to the right on this diagram, we see more downstream effects that will be impacted by pain and difficulty walking, but will also be impacted by many other things in life. So, for example, pain, difficulty bending knees, difficulty walking, may, of course, impact productivity at work, but so will a lot of other things. Walking or mobility aids that are available to the patient, the type of work they do, how far they have to walk at work, where they can park, what their job entails, whether the pain or walking difficulty has made them depressed so

it's mentally more challenging for them to overcome the challenges to get to work, the type of social support they have.

So it's difficult to interpret changes in these more downstream impacts like productivity without first understanding all of the components that may impact them. So also because of the added variability in these more distal concepts, it's more difficult to detect treatment change when measuring them within the context of a short clinical trial.

Thinking about clinical trials. From the regulatory perspective, evidence of treatment benefit must come from clinical trials that provide sufficient evidence to support claims of effectiveness and drug approval. And this evidence is obtained through the use of adequate and well controlled studies. And adequate and well controlled studies are based on a number of features, one of which is that the endpoints supporting efficacy are well defined and reliable.

So when is a clinical outcome assessment well defined and reliable; well, when there is empiric evidence to demonstrate that the score quantifies the concept of interest in the targeted concept context of use. And what does this mean, it means that we're measuring the right

thing in a defined population or targeted context of use, and that the score that quantifies that thing that we're measuring, does so accurately and reliably so that the effects seen on that outcome assessment can really be interpreted as a clear treatment benefit.

The PRO guidance describes good measurement principles that might be considered to evaluate whether a measurement is well defined and reliable. And all clinical outcome assessments, patient reported outcome assessments, patient reported outcome assessments, clinician reported, observer reported, and performance outcome measures can benefit from the good measurement principles described in this guidance.

Specifically, when we evaluate whether clinical outcome assessment is well defined and reliable, we evaluate the tool's measurement properties. So this includes content validity, construct validity, reliability, ability to detect change, and then information on how to interpret that change.

In addition, when we review clinical outcome assessments, regardless of the type, we need to think about all of the components listed here, the assessment, how it relates to the targeted claims, how it fits into the endpoint hierarchy, the conceptual framework, is there

evidence of content validity and other measurement properties, how do we interpret the scores, are there language, translation and cultural adaptation versions of the instrument and were these done using an appropriate process? What is the data collection method? Are we using electronic data capture that can help limit missing data? What is the clinical trial design and data analysis plan, how do these fit in and impact the measurement strategy, and how does the measurement strategy impact our trial design and analysis plan.

So when we think about clinical outcome assessments it's important to remember that assessments are supported by patients, clinicians and other observers and performance tests are not all clinical outcome assessments to evaluate treatment benefit in clinical trials. There are assessments that while reported by patients or other observers, are useful for very different purposes than as clinical trial outcome assessments. So these measures may be used for diagnostic purposes, prognostic purposes, used to select patients for participation in clinical trials, used for epidemiologic or population studies to better understand characteristics or the natural history of a condition, or used to assist in clinical practice decision making.

Assessments used for these other purposes are often not appropriate for use as outcome assessments in clinical trials, at least not without some modifications. So, for example, and instrument or measure might be great at broadly assessing improvements in an individual activity level for clinical practice decision making, so, for example, actigraphy, but that same assessment may not capture the right thing or concept of interest to inform a conclusion of treatment benefit in a clinical trial. It may not be sensitive enough to be able to detect changes, and it may not be able to support labeling claims that are not potentially false or misleading at the population level because we don't know exactly what is being measured.

So we do encourage early discussions with the FDA about outcome assessments. And there are two pathways to seek formal advice from the FDA. The first is through the traditional IND pathway for a particular drug development program. The second is within our drug development tool qualification process.

So the final guidance describing our drug development tool qualification process was published earlier this year in January, and while evidentiary standards are not described within this guidance, it does describe the process of qualification in detail.

So what is qualification; qualification is the conclusion that within the stated context of use the results of measurement can be relied upon to represent a specific concept of interest with a specific interpretation when used in drug development and regulatory decision making. So in plain language, within a specific clinical context, we're measuring the right thing in the right way, and we can rely upon the results of the qualified assessment across multiple clinical trials within that same clinical context.

Our resources are limited so we have to limit our qualification participation to only those assessments that are ultimately intended to support primary or secondary endpoints in clinical trials. And at the end of the qualification process, qualified instruments are meant to be publicly available. They're intended to be used across multiple drug development programs within the same context of use, so during instrument development we work with the instrument developers to insure there's broad applicability of the assessments under development.

And we have two fairly new communication tools on our website to help describe the elements to consider during the instrument development within the qualification program. The first is the roadmap to patient focused drug

development and then the second I'll discuss in a few minutes, is our revised wheel and spokes.

So first I'll describe the roadmap, which is intended to illustrate how we might embark upon a sound, orderly instrument selection or development pathway, beginning with the clinical context in which the instrument is intended to be used. So I'm going to walk through each of these individually, I know it's a little bit difficult to see the full thing on the screen here.

You see there are three columns, the first is understanding the disease or condition; the second is conceptualizing treatment benefit; and the third is selecting or developing the outcome measure. Often we see folks jumping right into column three, selecting an instrument, without giving any attention to columns one and two, to understand the elements of the disease or condition that may impact outcome assessment, and without fully conceptualizing treatment benefit before selecting an assessment to measure that treatment benefit.

When we think about clinical trial outcome assessment we have to do so by thinking about the concepts we plan to assess within a particular context of use, and the roadmap to patient focused outcome measurement can help us conceptualize a lot of these elements.

The roadmap is not intended to be yet another hurdle from the FDA. I know it looks very busy and there's a lot of information here, but, in fact, a lot of the things on here probably are already things that instrument developers and drug developers are doing, this just provides a suggested way to systematically think about these things to make sure that nothing is forgotten. And to think through these different elements that may impact your ability to assess and detect meaningful change in clinical trials.

So I will walk through each of these elements but first I do want to note here that this overview slide, column two is a little bit different than the version we have on our website because we are trying to make this a little bit more applicable to non-PRO clinical outcome assessments. And so I'll talk about that in a few minutes.

So first, understanding the disease or condition. We have to understand the natural history of the condition. It might be onset, duration, resolution, impact or conceptualization of treatment benefit and ultimately our choice of outcome assessments. So we think about the diagnosis, the pathophysiology, the range of manifestations, again, ultimately these might all impact how we think about measuring treatment benefit.

So consider patient subpopulations. Those listed here might not necessarily apply to all conditions, but are listed as checkpoints for those enlisted in research to think through which may apply. And there may be many other ways to divide groups into subpopulations.

It is important if there are expected variations in experiences in patients across different subpopulations that these are considered when selecting or developing outcome assessments. So in the case of pain and physical function, subpopulations may be based on location or type of pain, chronic versus acute pain, different severity levels of pain, leading to groups with different levels of physical function impacts, and these all might need different assessments.

Understanding how physical performance is impacted differently across patients with lower limb pain versus upper limb pain, will impact the type of physical function assessment that we should find appropriate in this context. So we wouldn't want to focus our physical function measure on walking ability in patients who are only having hand or wrist pain.

We might think about other subpopulations, including important patient phenotypes that could contribute to heterogeneity in the outcome, variations in

onset of symptoms or signs, daily variation in symptoms, and other aspects of biologic variability that could impact our outcomes.

So we also need to understand the health care environment. We want to identify currently available treatment alternatives and think about how that will influence our clinical trial entry criteria and design, identify clinical practice variations that may impact treatment, study design and outcome measurement. As an example, all patients with a COPD exacerbation are hospitalized in Spain, but not in the US. So if we're using COPD exacerbation or if we're using hospitalization as an outcome measure, we have to interpret that very differently in Spain than in the US.

It's also important to gather the patient and caregiver perspectives, what is their definition of treatment benefit, how do patients think about symptom or function burden, what's the most important from their perspective, and what should we incorporate into our outcome assessments? What is the impact of disease on life? Are there accommodations that patients are making to deal with their disease, symptoms, or impacts that we need to account for in our measurement strategy?

Patient or consumer input is very important. So I

want to take just a minute to quickly describe an initiative that highlights and is part of FDA's broader focus on continuing to seek patient input and encourage the involvement of patients in all aspects of drug development.

And this is the Patient Focused Drug Development Initiative. This includes systematically gathering patients' perspectives on their condition and available therapies to treat their condition. And we're holding a series of 20 public meetings over the course of 5 years and each meeting is focused on a specific disease area. So while we can't do a meeting for every existing condition, we do hope other efforts will build on the work that FDA is doing through these meetings and continue to engage patients when making drug development including outcome assessment decisions.

So column two, conceptualizing treatment benefit. And this is where the diagram I'm showing here today differs a little from our website version. We have tried to make this one, again, more applicable to non-PRO assessments, as well as PRO assessments. With non-PRO assessments you're often not, what you're measuring may not be the meaningful benefit, but some representation of it.

So we've split one box into two. And at the top of column two, conceptualizing treatment benefit, you see a

suggestion to identify the meaningful health aspect. That is the intended benefit to the patients in their daily lives. So it might be walking ability in daily life, for example.

The second box in column two then suggests we think about what exactly are we going to measure to conclude improved walking ability in our particular context of use. So it might be that we use a PRO assessment to directly ask patients about their walking ability in daily life, or it may be that we consider a concept like walking speed and performance in the clinic where we would use a performance test like 6 minute walk test to represent walking ability in daily life.

And the third box in column two reminds us to define the specific clinical trial context of use. And this conceptualization is based on the components of our understanding from having thought through all of the elements in column one. We've struggled with where to put this in the column because it really is an iterative process to select the meaningful health aspect, the concept of interest and the context of use. You can't determine what to measure unless you know what population you're thinking about, but you don't know where your outcome assessment fits into your trial objectives in the endpoint

hierarchy until you've selected your outcome assessments.

So elements of the context of use are considered before, during and after determining what to measure and how to measure it. The context of use and concept of interest decisions are iterative and done in parallel.

So we've tried to put together a list of common elements of the context of use that might impact concept of interest decisions. This list is not perfectly comprehensive, nor will every element apply to all drug development programs. But it is useful to give some thought to these different elements when making your decisions for a particular clinical program. So consider the disease definition, the patient subpopulations, clinical trial design and objectives, and the clinical practice and study settings.

Within the study design and objectives heading, we include the bullets endpoint definition and endpoint positioning. This is important in the regulatory setting and can impact our choice of outcome assessments to support endpoints, as well as the level of evidence needed to support the selection of an outcome measure. And note that outcome assessments and endpoints are not synonymous, but the score of an outcome assessment is used to develop an endpoint definition.

We also think about the following categories in the hierarchy of endpoints, so primary, secondary, and exploratory. For primary endpoints, meant to support a drug approval decision, a higher level of evidence is needed to support the selection or development of a particular outcome assessment that will form the basis of the primary endpoint and the indication statement and labeling.

Secondary endpoints are generally meant to support the findings from the primary endpoint. They may help us better interpret the primary endpoint or to learn and be able to communicate more about the drug and labeling. These assessments still need to have appropriate attention as they could be the basis for labeling claims and, therefore, these assessments have to be valid, reliable and interpretable.

Exploratory endpoints might be hypothesis generating, they might be used as additional supportive evidence to interpret the findings from the primary or secondary endpoints, but these assessments, the assessments supporting exploratory endpoints will not be the basis of labeling claims so these assessments might not need the same level of evidence and documentation to justify their use in clinical trials.

We talked about this earlier. Remember, the key is to define what is meaningful treatment benefit, what it is you wish to assess in the clinical trial, and then select the most appropriate clinical outcome assessment type and reporter given your particular study and context of use.

So when selecting or developing the outcome measure, once we understand the disease or condition and we've conceptualized treatment benefit, we can think about selecting and existing instrument, modifying and existing instrument, or whether there is a need to develop a new instrument.

The roadmap describes at a high level what steps to take in order to select or develop an instrument with the goal to use to support labeling claims. The first, we need to document content validity, then evaluate cross-sectional measurement properties and create a user manual. And I'll touch on these again when I talk about the second diagram, our wheel and spokes.

After the steps described in the previous slide, we move on to evaluating longitudinal measurement properties as well as considering how to interpret change on the instrument. Again, we'll consider this a little bit more in the following slides.

So now we'll look at our second diagram, the clinical outcome assessment wheel and spokes, and this is a revised wheel and spokes from the one found in the PRO guidance, and this one is meant to help describe the steps of instrument development within the qualification program. But again, these principles are applicable to instrument development for individual drug development programs, as well.

And there is a lot of overlap between our roadmap column two and three and the wheel and spokes, and this is just a high level of the wheel and spokes and I'll walk through each of the spokes in just a minute. And the wheel and spokes provides a bit more granularity in the specific things that we would review at various points in instrument development. And also describes the point in time at which qualification of instruments may occur.

So let's look at each spoke individually. So spoke one, identify the context of use and the concept of interest. We would outline hypothesized concepts and potential claims, determine the intended population, determine the intended application characteristics, perform literature and expert review, develop a hypothesized conceptual framework, position the clinical outcome assessment within a preliminary endpoint model, and then,

of course, document our decisions.

Spoke two is where we would draft the instrument and evaluate content validity. So obtain patient or other reporter input, generate new items, select the recall period, response actions and forma, select your mode or method of administration and data collection. Conduct cognitive interviewing, pilot test the draft instrument, finalize the instrument content, format and scoring rule and document content validity.

So what is content validity. It's extremely important and it's the extent to which the content of an instrument represents important aspects of a given concept for an intended use and for a defined target population. And I'm not going to spend a lot of time on this because our next presentation, Dr. Elektra Papadopoulos is going to talk in more detail about content validity and provide some good examples and describe some common pitfalls that we've seen.

So one way to help think about content validity though is to look at outcome assessments conceptual framework. A conceptual framework is an explicit description or diagram of the relationships between the questionnaire or the items in an assessment and the concepts measured. It describes how the individual items

contribute to the score that will be analyzed and ultimately described in labeling.

So here is a generic diagram of a conceptual framework, and you can see how the individual components on the left are used to support individual domain scores in the center and then those domain scores are rolled up into the total score that represents the overall concept.

So for physical function the overall concept might be physical function. If we're going to evaluate physical function and label and improvement in this, we use the conceptual framework to understand and insure that all of the relevant components that make up physical function in our context of use have been assessed in a comprehensive, well defined and reliable way.

So we ask ourselves, based on the evidence, are all important components of physical function assessed and included in our total score to justify a claim related to physical function so that the claim is not false or misleading? If not, can we clearly describe exactly what has been assessed to avoid false and misleading claims, and/or will another endpoint cover what is missing from our instrument?

So, for example, if we're interested in physical function as it relates to pain in lower limbs, we would

look at our item content to insure that it covers lower limb functions that are important to patients, maybe walking or climbing steps. But if we're interested in physical function that relates to all over bodily pain, more generally, we may look for domains representing lower body physical function, walking, as well as upper body physical function. So use of hand for fine or even gross motor tasks.

Now spoke three is begun after content validity is established in our context of use. This is where the final version of the instrument with its scoring rule can be tested for reliability and cross-sectional evaluation of construct validity, administration procedures and training materials are established, and a user manual is prepared. And, of course, all of this is documented.

And at this point, CDER can consider qualifying this instrument for use in exploratory studies. This means that we agree with the content of an instrument and believe it is measuring what it set out to measure. However, we don't yet have a full understanding of the instrument's ability to detect change and how to interpret meaningful change. So, therefore, we are recommending the instrument be made publicly available, used in exploratory studies, typically phase II trials, in order to have the needed

information to plan for the use of the instrument in phase III confirmatory trials.

At this point, if the instrument is used in an adequate and well controlled confirmatory study as a primary or secondary endpoint and the clinical trial demonstrates a treatment effect that is clinically meaningful, it might be considered for use to support labeling claims. Wile the assessment is not yet qualified for this use yet, it may be implemented in this way with some additional risk to the drug developer given that there's a lack of information available about the longitudinal measurement properties and the instrument's ability to detect and interpret meaningful change. So we would encourage discussions about the use of this assessment in this way with the appropriate review division in the agency.

After we're comfortable with content validity, in particular, we can evaluate other longitudinal measurement properties described in spoke IV, so again, assessing ability to detect change, longitudinal construct validity, identify responder definitions, provide guidelines for interpretation of treatment benefit and relationship to the claims. And document all those results, also updating our user manual. And now, after the successful evaluation of

these longitudinal measurement properties, we can consider qualifying the assessment for use in adequate and well controlled studies as a primary or secondary endpoint.

After an instrument is qualified for a particular context of use, the assessment might be considered for other contexts of use. In some cases we may need to modify the instrument for the new context of use, so spoke V is included in the wheel and spokes to account for this potential.

And in conclusion, the roadmap a well defined and reliable outcome assessment begins with a full understanding of the disease or condition to be tested. An assessment cannot be chosen or developed without a well defined context of use, understanding of the meaningful health aspect, and targeted concept of interest to be assessed, the science of measurement is continuing to evolve with new tools and methods for efficient development and modification of assessments. And there is no one size fits all approach to measurement development.

So we have tried to organize some ways to think about measure development and suggestions on good measurement principles in our PRO guidance, as well as other communication tools described here today. However, we all need to remain flexible while we implement the best

instrument development decisions for a particular situation, population or drug development program, and we encourage instrument developers and drug developers to talk to FDA as early as possible when planning for drug or instrument development. Thank you.

TURK: Well we're going to save questions until the panel so we can have an opportunity for you to think about your questions to present them to the entire group. That's a tremendously useful presentation and although this was being presented from the FDA perspective for drug development, I would suggest that 98 percent of what was presented would be just as relevant for looking at physical therapy intervention or any other non-pharmacological intervention. But that the concepts, and spelled out so clearly, are extremely helpful to us.

Now a lot of information was presented, and I should have mentioned that in the past what we have always try to do at IMMPACT meetings is with the permission of the speakers, get access to their slides and put them on the IMMPACT website so that you'll be able to look at these. Because a huge amount of their useful information but it goes by fast, and that's going to be true for many of the other people. So I'll ask each of the speakers' permission, if they give permission we will put those on the IMMPACT

website for you to have access to. If any of the speakers, by the way, have any slides that they are proprietary, they don't want, we'll ask you could you remove that and is that permissible to put up there. So we'll do the best to get to that information, but thank you so much, it's a tremendous beginning, but I really want us to, I want to think about that 98, maybe 99 percent of what was presented is not specific just to drug development but good practice in developing any clinical outcome for any clinical trial.

So let's move on now, and we'll save questions until the panel, to our next speaker, who will also give a perspective from the FDA on how they actually are looking at different measures that are being suggested as outcomes that might be used in clinical trials. And our next speaker is Dr. Elektra Papadopoulos, am I saying that correctly? One of the things you are going to learn about me is that if there is any way to screw up how to pronounce your name, I will do it, and if I say it correctly, it was in error, it should have been screwed up.

So Elektra, what you're going to be talking about is the FDA's approach to the review of outcome measures for drug approval. So we heard about the process in developing some of these measures and now we're going to hear about how is that information used by the FDA moving forward.

FDA Approach to Review of Outcome Measures for Drug Approval and Labeling

ELEKTRA PAPADOPOULOS: Thank you very much, I'm very happy to have this opportunity to talk to you today. I'm going to be talking about our approach to review of content validity. So these are my own opinions and do not represent and official FDA position.

I'll start by describing the elements of review of content validity and then in the second part I'll go into some of the common design shortcomings that we've observed over the years.

So this is the slide that Dr. Slagle has already presented, the wheel and spokes slide. She has presented a broader context of the importance of identifying the concept of interest and the context of use before setting out to identify or develop an instrument. And what I'll be doing, is I'll be focusing on spoke II of the wheel and spokes, which is drafting the instrument and evaluating content validity.

As we have already heard, content validity is the extent to which the content of an instrument represents the important aspects of the given concept in the context of use in a defined patient population. And this is informed by both qualitative and quantitative research. Importantly,

qualitative research from the targeted population of respondents is critical for insuring content validity.

So how do we think about content validity and what do we review? Essentially we review all of the elements of the instrument. This includes the user manual, the instructions, mode of administration, whether it's by self administration or interviewer administered, the data collection method, such as paper based or computer assisted, and its scoring algorithm. So all of these are essential.

Now, many sources of information come together to support content validity. Oftentimes a literature review is done at the very beginning. Expert input is critical and this includes clinical expert input and importantly also experts with experience in instrument development.

Once we've developed the hypothesized conceptual framework for the instrument, we then can, for a PRO instrument we would seek patient input and we would use concept elicitation interviews to generate item content, followed by cognitive interviews to evaluate patient understanding of the instrument.

And this table, I apologize, it's difficult to read, it's from part II of the ISPOR Task Force report published in *Value in Health* in 2011. The paper

recommends giving some thought, some early thought to criteria for development of the items. And this table includes many criteria such as does the item capture the concept it was intended to capture, is it relevant to all members of the target population. Is the item worded in a manner consistent with expressions used by patients. Does the item reflect different severity levels of magnitude of frequency or severity, and so on. And similar considerations are also used by the FDA when we're evaluating item content.

So now I'll switch gears and talk about some of the common design shortcomings in item development. This slide presents an overview of several of the common shortcomings. It's not all inclusive, but when we see these they definitely raise red flags for us. They include global ratings of complex concepts; ratings of change; assessment of unobservable concepts by an observer, such as when a parent or caregiver performs an assessment of the patient; measuring indirect impacts without also measuring the direct impacts, without also measuring the core signs and symptoms; an appropriate recall period; and, combining more than one concept within a single item.

So a global rating is usually a single item rating of a complex multi-domain concept. And in this

example we see the question: "Overall, based on all your years of medical practice, how severe is this patient's overall health condition?" This is a real problem for clinical trials. Respondents need to consider both known and unknown aspects of the patient's condition, and assign a single value or rating in an unstandardized manner. So the issue is that we're not clear what it is that is being measured and whether different clinicians are using similar criteria when assigning their ratings. So to avoid this problem, we would select or develop instruments that include separate items or scores for each important sub-concept of interest.

Items that ask respondents to rate change are problematic because they require the respondent to recall previous state, which is often long ago. The respondent must then make a comparison from their current state to the previous state, and the final value doesn't represent an absolute severity level but just a comparison. So for this reason we recommend items that describe the current state without requiring any comparison to a previous time point.

This is an example of, say, when a parent, teacher, or caregiver is asked to perform an assessment outside of a clinical setting. It's important, again, to show that the reporter is only reporting on aspects that

they could observe, such as through one of the senses. And because no one but the patient can actually know what that person is feeling, an observer cannot report on things that are unobservable.

So instead of asking about how severe is your child's pain, we would ask the rater to make observations on things like how frequently did your child cry, or did they wince or hold or guard a body part, for example.

Now as Ashley said before, when we're evaluating treatment benefit, we recommend the core disease defining concepts be assessed first and considered first before we think about more downstream effects. However, oftentimes existing instruments include phrasing that asks about more downstream effects. And in this case, the example is I'm afraid I won't have enough time to reach the bathroom or I worry my incontinence will get worse.

So these really are measuring emotions such as bother, distress, or being afraid, and do not measure the effect of treatment on the core disease symptoms. So in this case, for a treatment that's designed to treat incontinence, what we're really interested in is whether or not the number of incontinence episodes has decreased. And only then can we begin to interpret some of the more distal impacts of the disease.

An important consideration for content validity is that of recall period. The recall period should be appropriate for the patient population, concept of interest, and clinical trial context. In this case, is it reasonable for patients to remember how much itching they've had over the past four weeks? If itching varies day to day or hour to hour, how can a patient mentally average their itch severity over that lengthy period of time.

Also, if the trial is of short duration, the recall period may actually be longer than the treatment period. So while a lengthy recall period may not be necessarily a showstopper if a treatment effect is very pronounced, for more subtle effects, however, the variability of a lengthy recall period may attenuate the effect size enough that an effective treatment may be missed.

And this is an example of an instrument where the item includes more than one concept: "Today, how many times did you cough up blood or sputum?" In this case, it's impossible to disentangle which concept the respondent is responding to. So a better method would be to separate this into two items so that a frequency of coughing blood can be measured separately from the frequency of coughing

sputum.

Now to illustrate content validity or lack of content validity in this case, in relationship to a PRO instrument, I'd like to discuss a hypothetical example. In this example, the study population is primarily bedridden, and even minimal physical activity such as walking to the bathroom is very difficult. The clinical outcome assessment aims to measure physical functioning; however, included in the assessment is a question, do you have trouble running to the bus. So clearly, this is a case where the content of the assessment does not represent what we're interested in, in this particular target population. In another population, a more active population, this may be appropriate however.

Here's another example for a performance measure. Again, this is a hypothetical example. So in a population where the patients report mobility as their biggest concern, we do qualitative research and we find that these patients find that when they, their biggest problem is not walking, but the ability to get up and out of the chair. So a performance test in this case, such as a 6 minute walk test that assesses gait speed at which patients can walk at a clinic, is not measuring the relevant component of mobility in this patient population.

So in a patient population, speed of walking and ability to get out of the chair are both important. Therefore, only measuring walking speed using the 6 minute walk test, does not provide a comprehensive assessment of mobility in this patient population. So this really exemplifies how the content of the instrument needs to be considered in relation to the context of use in the defined patient population.

And I'll show one more example of the importance of matching the instrument to the patient population. This slide shows the physical function domain of the commonly widely used health status measure, the SF-36. This domain include ten items and you can see more than half of these involve general activities or specifically refer to primarily lower limb activities such as climbing stairs, bending, kneeling and walking.

Of the ten, there are only two that involve upper limb functions, carrying groceries or bathing and dressing. So we can see from this that the physical functioning domain content may be more appropriate for measuring lower limb functioning than for measuring impairments in upper limbs.

This graph shows the mean change in the SF-36 physical summary score, 12 weeks, following 20 different

therapeutic procedures within a network of British hospitals. I'd like to draw your attention to the two points on the graph. One is carpal tunnel release, and the other is hip joint replacement. And for this we can see that clearly the physical function summary score is a much more sensitive indicator in the hip joint replacement population, as this is primarily a lower limb activity population.

Now I'd like to say a few words about the use of PRO measures to assess performance of daily activities. Often instruments ask patients whether or not they're able to perform a certain activity but may actually be leading patients to respond on the basis of their desired condition rather than on their actual condition.

The PRO Guidance provides specific advice to avoid this. It says, "in assessing the concept ability to perform daily activities, it is more appropriate to ask whether or not the patient performed specific activities an dif so, with how much difficulty, than whether or not the patient perceived they can perform daily activities, because patients may report they are able to perform a task even when they never do the task."

Now Ashley mentioned actigraphy in her talk, and I'd also like to say a few words about it. When we're

thinking about possibly incorporating actigraphy measures in clinical trials as key endpoint measures, we need to step back and ask several questions first. We understand physical activity is a very important measurement goal in situations such as chronic pain, but importantly we need to understand what aspect of physical activity is being measured by the actigraphy, whether the results represent a meaningful assessment of activity, and what kind of data would be capable of supporting the use of actigraphy to substantiate physical functioning claims.

Importantly, there are some, we need to even match the context of use to the actigraphy. So, for example, in patients with lower back pain, frequently they're limited in aspects of functioning that may not be captured by actigraphy such as the ability to stand in one place for a long time, or lift and carry objects. Again, we need to consider the adequacy of the clinical outcome assessment in the context of the specific clinical trial population.

So to summarize, content validity is a critical measurement property of all clinical outcome assessments. Without content validity, it's hard to detect a treatment effect oftentimes, and if a treatment effect is detected, it is then often difficult to accurately describe the

benefit in labeling.

Careful attention to content validity and of instrument design is critical to insure that the outcome assessment can be interpreted and described in product labeling in a way that is accurate and not misleading.

Thank you.

Q&A and Panel Discussion on FDA Approach to Outcome Measures

DWORKIN: Ashley, could you come up, and also Dr. Rappaport. For those of you who haven't yet met Dr. Bob Rappaport, he, I think as you already heard, is the Director of the Division of Anesthesia, Analgesia and Addiction Products at the US Food and Drug Administration, and as Dennis mentioned earlier, Bob really was the person who had the idea of and spearheaded the ACTTION Public-Private Partnership, and I also think it's safe to say that without his support over the years we wouldn't be here at this meeting. And so we appreciate Bob's commitment to the development of better pain treatments.

So this is really a period now, we have about 20, maybe 30 minutes before a coffee break, for you all to ask questions of the two speakers we've had this morning, and even more general questions with respect to FDA's views and perspectives on qualifying outcome measures. So would someone like to start the ball rolling? Bob.

ROBERT KERNS: Bob Kerns, Yale and VA Connecticut Health Care System. I'm curious, I'm surprised to learn that the FDA is in the business of measurement development, and I guess I'm thinking that the implications of these presentations is that it's okay or it's feasible to

actually combine measurement development with design and development and testing of a clinical intervention. I find it hard to imagine that the NIH would consider an application for a clinical trial or score it well if in the context of the design of the clinical trial, it was also trying to develop an outcome measure or do both things in the same, in the same application.

So I'm just curious about that idea, I would think that measurement would come first and clinical trial would come later.

SLAGLE: I think in an ideal world the measure development would come first and the clinical trial would come later, but I think practically speaking a lot of times that just doesn't happen. And so we're often in the position where we're developing measures during a drug development program.

So, you know, to the extent possible, doing as much measure development before phase II studies is important. In your phase II study then you can be collecting some of the longitudinal data that is useful to plan for your phase III study.

KERN: Well I hear that that's what you're doing, I am skeptical that that's possible to avoid confounder bias in that, in either process, frankly.

SLAGLE: And I just want to clarify, FDA is not in the business of measure development, so we will work with groups who are developing measures and so they're setting the timelines, but we provide advice and consultation and if they have questions they can certainly ask us questions through the individual drug development program or through the qualification process. But our role is to provide advice and consultation rather than actually --

KERN: So it seems like a bad idea.

DWORKIN: So Ashley, I think I have a follow-up related question, qualification of a measure, is demonstrating a treatment difference a necessary part of qualification. In other words, can my measure not be qualified until I've shown that it separates some treatment, you know, typically in a superiority design, from another treatment, or is that not necessary? It was unclear from the presentations whether discrimination, to use the OMERACT phrase, is required before a measure is fully qualified?

SLAGLE: To be fully qualified, yes, we need to have that information, that it's able to detect change, and that we can interpret that change. There is a point where we can qualify earlier in the process, when we don't have

that information yet, but we qualify the instrument to say we agree with the content, you've done the qualitative research, we think that this has good potential to be able to detect change but we don't know that yet. So there is a first qualification, the instrument would be publicly available, used in multiple settings where we can accumulate more information about ability to detect change, then it would be qualified fully --

DWORKIN: Okay, so I understood about the detect change, it wasn't exactly what I meant, I meant the ability to detect a group difference between an active treatment and placebo. So I've got to measure that shows change over time let's say with worsening, but do I also have to show that the measure can discriminate one treatment from a control treatment?

SLAGLE: That is not part of the qualification process simply because I think that you would need to have an effective treatment first in order to detect that. So sometimes we don't have that and --

DWORKIN: So qualification can occur with simply the ability to detect change, interpret change, but not a group difference in a clinical trial?

SLAGLE: That's right.

DWORKIN: Dr. Lee Simon has a question.

LEE SIMON: I was wondering --

DWORKIN: Where are you from, please?

SIMON: Lee Simon, Boston. I was wondering, many in rheumatology are seduced by the idea that a patient global response is a very meaningful and descriptive example of how a patient feels in a responsive measure. You didn't talk about what could be considered as a primary outcome in the context for approval, versus something that might be informative for the stakeholder to understand potential benefits. You alluded to these global possibility of outcomes and suggested that they're too complex to be specifically useful to understand responsiveness.

So could you comment on the utility of the patient global in how you think about these issues and whether or not A) it can be used, which it never has been that I'm aware of, for a primary outcome, and also where you use such a thing, if at all, in understanding responsiveness?

SLAGLE: Well global outcomes, as I mentioned, do have concerns because if it's a particularly broad or multi-domain concept, it's impossible to tell which of the different domains is actually changing with treatment. That said, oftentimes global assessments are used as

exploratory endpoints and in that context they can actually be used as an anchor for interpreting changes on multi-item instruments. So they do have utility in that regard. But as far as trying to describe what the treatment is actually doing, they've very limited in that.

BOB RAPPAPORT: We do use them in the pain trials as secondary outcomes along with some others, as we usually ask people to do a global as a secondary to get a sense whether there is going to be, just how consistent things are. So if the primary, which is almost always pain, is positive, but then you've got a global that's going in the opposite direction, that tells us there's a problem with they study.

DWORKIN: Could I just extend that one more second? In thinking about that, and I was particularly stimulated by Dennis' introduction, if pain is the primary outcome, some manifestation of pain response, he suggested that the interpretability of how much pain did you have in the last 24 hours or what was the average pain you had in the last 24 hours, that kind of question, is that an acceptable context in that patients will have a different interpretation of what average means, and your instructions may be quite obscure because our anticipation of what a patient might feel may be quite different than what the

patient interprets on how they feel.

So at the present time, what's the present standard for what you're thinking about in the context of asking the question of what kind of pain they're having, is it right now at the time you're asking the question, a recall question of 24 hours which may or may not be particularly applicable as you've mentioned, how are you now thinking about this particular question of your primary outcome of a pain response?

RAPPAPORT: I think there's a certain amount of variability depending on the study and I'm actually going to ask Sharon to respond to this because she's a little bit closer to what we're doing with this.

SHARON HERTZ: I am Sharon Hertz, I'm Deputy Director in Bob's division at FDA, and Lee, I can't help but wonder if perhaps there is something behind your question (laughter).

SIMON: Never. I never have any other purposes.

DWORKIN: It was a little off target.

SIMON: No, it wasn't, it was brought up already.

HERTZ: A lot of our use of globals are sort of remnants of the past, we inherited this in the context of what was done in rheumatology for a while in the signs and symptoms of osteoarthritis because that indication was

based on three primary endpoints, pain, function and global. We found this to be over time not the most useful set of primary endpoints for a couple of reasons, but I don't think I need to necessarily go into that right now. But the global is sort of a carryover from that, it's still in the signs and symptoms of, actually it's in the -- it's still attached to fibromyalgia, the syndrome. So it's not a especially informative really because it generally tracks with pain, as does function, and sometimes it reaches statistical significance, well many times it reaches it, and sometimes it simply doesn't, even though pain may have, similarly function may or may not have reached statistical significance even though it will trend pretty much with the others.

So we use them, I don't know that we really find them particularly useful.

RAPPAPORT: Sharon, in regard to Lee's question specifically regarding what question we ask for the primary outcome, is it average pain over, I know it varies from some areas, some study areas like acute pain versus chronic --

HERTZ: So the primary pain we don't require a specific one, and we ask, we recommend worse pain in 24 or 48 hours because we feel that reflects something patients

can accurately convey. And we often accept average pain because for some reason investigators feel quite wedded to it. And it probably more likely reflects worst pain and we recognize that, so even though it's called average pain and we realize based on our work with SEALD that it probably is more reflective of worse pain. It's still pain and it still seems to be relatively reliable and we are willing to accept it although we often request that we use what's actually being measured instead of what's not quite being measured.

DWORKIN: Okay, there are a lot of questions. John.

FARRAR: So two things, one very quick, and the second one I think a little longer. The first one is that I agree completely with Bob Rappaport which is that the purpose of a global as an outcome is to make sure that everything goes in the right direction. The patient's pain might get better but their nausea is so bad that they can't take the medication. And so it gives us another indicator of making sure that things go in the right direction and it doesn't necessarily need to reach statistical significance to still go in the right direction. So I think those are both important.

I have to say that with regards to worse pain,

what I come back to, which is really what was presented in both the first and the second talk, which is you've got to ask the right question. And the issue with regards to the measurement of pain is that worst pain certainly can be the right question in certain circumstances, but there are multiple circumstances where worst pain is not the right question. And I would argue that in an arthritis study, a patient who gets up and walks to the store and has absolutely terrible pain might remember that worst pain and report it reasonably accurately, but, in fact, they're much better on their drug because their worst pain doesn't last as long, it's on average not bad, and so it really I think depends.

If you're trying to treat that worst pain episode, then for sure, worst pain is the right measure. If you are trying to treat their overall pain then I think that the average pain over the course of a day is a reasonable measure when measured 7 days in row with an average.

The last piece of this is that we need to admit that pain is subjective and that there isn't any way of getting around it.

DWORKIN: Sharon.

HERTZ: But John, the problem is when we ask

people for average pain, are we actually getting average pain? So I agree with you, the question you ask is extremely important, but my understanding is that when you ask people to average their pain, what they're remembering is more a reflection of their worst pain. And that's why I say we should ask what we're getting for. And I don't know how, what does that mean when somebody is trying to average that horrible walk to the grocery store versus feeling good sitting home, what does that average pain mean? If it's somewhere between 1 and 8 and they say it's 4, I mean what does that mean if it goes to a 3-1/2 or if it goes up to a 6, does it mean they were feeling better so they're walking more?

So I think if we really want to know what we're measuring, the question is what's your pain right now? And then if you get multiple --

DWORKIN: What I would like to propose, because I think we could easily hijack this meeting right now and spend the next day and a half talking about the conversation that is happening back and forth between Sharon and John, so what I would like to propose is that we add how do we assess pain as a primary endpoint in clinical trials to our list of possible topics for next year's IMMPACT meeting, where we can devote two full days to the

conversation that just occurred between John and Sharon. We have a list that we can come back to, as Dennis said, tomorrow afternoon, and I am adding this discussion to the list of possibilities for next year, because obviously it's a critically important question, current pain, worse pain, average pain, how do you assess it, how many times a day. And, you know, the last time we talked about this at an IMMPACT meeting, I don't know about OMERACT, the last time we talked about it at an IMMPACT meeting was in 2003. So maybe it's worth revisiting after a dozen years.

So Ajay. You question has to be about either FDA qualification or physical function, it can't be about pain (laughter).

AJAY WASAN: So I'm Ajay Wasan from the University of Pittsburgh, and so I just want to take a step back, so just get some clarity and perspective. So I was really struck by the comments that scales that are involved, you know, ratings have changed, global ratings or recall periods are problematic for different reasons. But at the same time for many of those measures the measurement science on them is excellent showing an incredible content validity, reliability, reproducibility, sensitivity to change, everything you would want in a measure. So it just seems to me at odds with the measurement science or the

perspective you're presenting. So I just wonder if you could reconcile that a little bit for me and clarify how the recommendations fit in with what all the other measurement science and the psychometrics of all these measures tell us.

PAPADOPOULOS: The focus of my talk was on content validity, which is something that comes before assessment of all the other measurement properties, we test reliability, construct validity and ability to detect change. And so our position is if you don't first have content validity, all those other measurement properties aren't particularly meaningful. So the criticality really of content validity is that we need that in order to describe what is being measured in a way that's accurate and not misleading to patients and providers.

SLAGLE: I also think, we want to make the point that content validity is specific to the population and we also think about it in terms of the clinical trial context of use. And so in some cases a lot of these measures are very well understood and are perfectly appropriate for other settings but in the context of the clinical trial then some of these elements cause particular problems and some pitfalls in clinical trials.

So no instrument is necessarily validated for all

uses, so we have to look at the measurement properties and where they were established and whether they're applicable to the clinical trial context.

DWORKIN: Bob.

RAPPAPORT: So I would like to go back just to Bob Kern's question because I think that got kind of, it's just hanging out there. Could you expand a little on why you have that concern or what exactly it is that you're concerned about so we can address it?

KERN: Yeah, so I guess to me in thinking about the design of a clinical trial, decisions about measures are critically important, they're central to the design and methods of the study. And so I guess I was imagining that it would be hard to think about how one would both be focused on testing the benefit of and the outcomes of a new intervention or drug, novel compound, and be developing the measure at the same time. It seems that it's important to set the bounds of the design of the trial and the testing of the new intervention and so forth and all that based on decisions that are made on the best that you have at the outset in terms of measures of potential outcomes.

In terms of FDA being involved in, and maybe it actually even speaks to Ajay's question, I've always thought about measurement development as being in the

domain of measurement science, if you will, right, and not I guess mixed up in the idea of people that are doing clinical trials, having different sciences. And so these presentations seem to imply a new way of thinking to me that I was surprised to hear about which is that they go hand in hand or could.

RAPPAPORT: I'm going to respond and then I'm going to ask my expert colleagues who really know about this stuff to respond as well. But a couple of things. One is I think what was said was that these qualifications should occur during phase II where you are doing a lot of work that's still exploratory. When we get into phase III trials we're not qualifying at the same time, and when we're looking to define whether a drug is actually effective in a phase III trial that's going to be used to determine whether the drug can be marketed, we have to have a qualified tool at that point.

So this, and the other thing is that I think that what, and you can confirm this for me, what the FDA is doing, is trying to provide help and guidance in qualifying tools so that when we get them in phase III they're really good. We don't do, as Ashley said, we don't do the actual qualification, we just provide guidance as to what we will be looking for when it's used in phase III.

KERNS: So maybe that was the first point that I made, I didn't know that the government was in, or the FDA was in the business of measurement development or advising people about that. But that always seemed to me separate and outside that. And in terms of measurement development, I've developed measures as many know, and I've always been clear in discussion sections that this remains in the research domain despite efforts, you know, people wanting to rush to put it into clinical settings or clinical context. It takes years for measures to kind of meet some accepted standard in the field, to be credible measures, to start to converge on construct validity, which is never achieved, by the way, in my view, it's just converging on construct validity. And I'm not sure that, I guess I'm wondering about the FDA's, the FDA ultimately has to make some judgment about whether the body of evidence supports that, but I don't think that it's something that happens in some short-term period, even in the context of up to phase II trials of a specific drug. I think I'd be reluctant to encourage the idea that that's even feasible until a new measure and new approach to measuring some important outcome actually has stood the test of time so to speak.

SLAGLE: I want to make one point in that I think part of the problem in the past is that measurement

scientists and people developing clinical trials haven't worked together and so we have a lot of measures that have been developed outside of clinical trial use that don't work at all in clinical trials. And so having a situation where both expertises are coming together is actually beneficial. And the FDA does, we do provide a lot of responses and advice on instrument development. Our group, Study Endpoints and Labeling Development, we consult to all of the divisions in the Office of New Drugs on just that. So we are very involved.

The other thing is, and we are criticized for this often, is that it does take a long time to develop new measures. And so we are looking for an accumulation of evidence, and that doest take quite a long time in many cases, so I completely agree with you.

DWORKIN: We have to have a coffee break soon. Dan Clauw, you've been waiting patiently.

DANIEL CLAUW: Dan Clauw, University of Michigan. I just have a question about sort of qualifying a domain rather than a specific outcome measure. I think it's becoming increasingly clear in a number of rheumatic diseases that symptoms like fatigue are often just as impactful on function, negatively impactful on function, as symptoms like pain. Susan Murphy's done some really

interesting work in the OA cohorts, for example, and obviously we know that's a big issue in conditions like fibromyalgia.

So I'm just wondering if there would be a comparable regulatory pathway to get something like fatigue to be a primary outcome, because I think what we're learning, especially in drug development, is that the drugs that would improve fatigue are often quite different than the drugs that would improve pain. And we shouldn't really assume that all the drugs that might help improve fatigue and then might improve the function of our patients with rheumatic diseases are also going to be effective analgesics and that we're going to be able to get that pain primary.

So I'm just wondering what the FDA's thinking is about qualifying a domain rather than a specific outcome measure?

SLAGLE: When we think about qualifying assessments or domains like fatigue, we have to think about it for what context of use. So we do think about fatigue in particular patient populations. If there is evidence to support that a particular assessment of fatigue is appropriate in that patient population, then of course that's open for qualification. But we really, we are

seeing across multiple conditions that fatigue is very important, and so as we get more experience with fatigue in individual populations, there may come a time where fatigue is something that can be thought about across many different conditions, but we have to have the evidence to -- we have to build on the evidence to understand that first.

CLAUW: But again, just for clarity, what kind of evidence can we help collect that would then convince you that this would be an appropriate primary in some conditions?

RAPPAPORT: Well I don't think you have to convince us that it's an appropriate primary in some conditions, we already believe and, in particular in fibromyalgia, that fatigue is one of the major things. And it was at, we had one of the patient focused drug development initiative meetings recently, just a few weeks ago, with the fibromyalgia community and we heard from a lot of the patients that fatigue is really, you know, pain is pretty bad but the fatigue is really what's driving them crazy.

So, you know, we're looking for a tool to measure that so that it could be used as a primary outcome. If that's what, you know, the drug treats, I have no problem

approving it for fatigue in fibromyalgia patients. I just need to know how to measure it and that's where I need help from the SEALD folks.

DWORKIN: Okay, so, Jim, it sounds like your question is directly related to this, so Jim and then Kristine, you had a question I think? And then we'll have coffee, so three more questions then coffee.

JAMES WITTER: Thanks, Bob, Jim Witter from NIAMS and the PROMIS initiative, not a question, just a comment that we are working with FDA to qualify the domain of fatigue. As we speak there are ongoing efforts, so I'm happy to hear this conversation.

DWORKIN: Kristine.

PHILLIPS: You mentioned the common shortcomings that are all patient reported outcomes, can you also comment on the common shortcomings for performance based measures of physical function, like 6 minute walk?

PAPADOPOULOS: I personally don't have a great deal of experience yet reviewing performance outcomes or qualifying performance outcomes. Most of the examples I've provided were taken from, as you noted, reviews of patient reported outcomes or clinician reported outcomes.

PHILLIPS: Could you comment on the levels of types of evidence that might be needed in actigraphy

performance outcomes, for example, I think that would be helpful. The kinds of things that FDA needs to review to determine whether actigraphy is measuring the thing that we would put in the label.

PAPADOPOULOS: So that's really, I think of it as a two-step process. You know, one is what is the immediate thing that this device is measuring. We need to insure that the device is actually accurately measuring what it says it's measuring, whether it's acceleration or something else. And so that's where we rely on our colleagues in the Center for Devices to insure that it meets those specifications.

After that step, we also then need to link that measurement, those acceleration data, to something that is actually meaningful to patients in their daily lives. And so we do need evidence in order to be able to infer something about treatment benefit. And I guess the example I provided, just a hypothetical example, was, you know, if someone is really impaired in their ability to lift or stand in one place, that is going to be completely missed by using an accelerometer.

So those are some of the things that we're thinking about as we're now engaging in these reviews for performance measures. It's challenging.

SLAGLE: I think one of the common problems with performance measures is that it's something that is considered more objective and easier to measure, but we don't always know exactly what aspect of function it's measuring. And so it's a real challenge to link the performance measure to what's the meaningful benefit and to be able to describe that benefit. Just because we can measure something well doesn't mean it's the right thing to measure, that we even know what's being measured.

RAPPAPORT: I think, this isn't exactly a performance measure, but I think another example where very well accepted hard endpoint measures can be very difficult to interpret is polysomnography when you are looking at sleep outcomes. I mean there's so many endpoints you could look at there and which of them are important to which patient population. You know, we started to ask the fibromyalgia studies to look at polysomnography so that we have, because sleep is actually a huge thing for that population, but still we're not sure which endpoints to look at and how to interpret them even if we could tell people exactly what endpoints to look at, it's a really challenging area.

PAPADOPOULOS: Can I mention one more thing about, you know, how do we think about these performance

measures. I do think we need to gain input from patients, themselves, as far as how closely these measures approximate some of the difficulties that they experience in their daily lives. And so I do think qualitative research would play a valuable role in the evaluation of performance measures, just as it does in patient reported outcomes.

DWORKIN: So clearly a recurring theme for the next day and a half is going to be the pros and cons of patient reported approaches to assessing physical function versus performance and clinician observer measures. So this is an issue we're going to be coming back to again and again and again I think.

And so why don't we have one last question from Ian and break for coffee.

IAN GILRON: Ian Gilron from Queens University in Canada, thank you for excellent talks. My question has to do with knowledge synthesis of empirical evidence of validity and reliability of outcome measures. It's not an FDA specific question but thanks to the meeting organizers we had some very thought provoking readings coming up to the meeting. I'm just wondering whether, there have been a lot of groups that look at study characteristics for treatment efficacy for example in terms of what's the ideal

sample size and trial duration for studies to be valid in systematic review and meta-analysis. I'm just wondering have there been any consensus recommendations on study characteristics for validity and reliability of outcome measures to, you know, sort of looking at some of the systematic reviews that we looked at, some of the evidence was based on considerably small studies, and I'm just wondering whether that fits into the acceptability of certain outcome measures.

SLAGLE: As far as the consensus recommendations for outcome measure development, there have been a series of publications by the International Society of Pharmacoeconomic and Outcomes Research, ISPOR, that describe, you know, good principles in instrument development. I referred to the publication in 2011 by Patrick, et al, and that one was focused on PROs but there is also a working group or a task force convened that is discussing recommendations for development of clinician reported outcome measures. And so we're expecting a publication later this year on that topic.

And I think, you know, the patient reported outcome guidance is a very useful resource, as well. a lot of graduate schools use that document in teaching their students about instrument development. So it has really

some very good information, not only about how we review them, but how they are developed, as well.

DWORKIN: Dorcas, isn't it the case that COSMIN also provides guidelines for how to evaluate studies that are, to develop new measures?

SLAGLE: Yes, so there is the international consensus group that's based out of the Netherland, the COSMIN group, that is trying to establish some criteria for the quality of the studies. And then there is also MPRO, which is out of the Medical Outcomes Trust, the scientific advisory committee of the Medical Outcomes Trust.

There's a number of different appraisal systems that are coming up, we found about 11 of them now in the literature to sort of try to help us evaluate the quality of a given study of reliability or of a given study of validity.

DWORKIN: And in the publication from this meeting we should certainly mention those approaches. Okay, so it is 10:10, let's have a 20 minute coffee break and reconvene here at 10:30. Thank you all.

(off the record - break taken)

Conceptual Overview: What Exactly Do We Mean by Physical Function

TURK: Once everybody gets in the room we'll get started back up again, I'm glad to see there's so much discussion and dialog and debate going on. I should mention while people are coming in that tomorrow afternoon after the formal presentations we have about a three hour block of time in which we will start tying to synthesize, pull things back together to help us move towards the manuscript that you all will be seeing.

As soon as everybody is in the room we're going to get started back up. There is one clarification that Dr. Rappaport asked to clarify something, so let me just wait till everybody is sitting down, Bob, before we do that and before we get into formal presentations. And this was in relation to the morning, the beginning prior to break section, Dr. Rappaport wanted to clarify a point. So Bob, I think 90 percent --

RAPPAPORT: Okay. I just want to make it clear that, and Ashley and Elektra, if you want to add to this please do, but the agency isn't saying that old tools need to be requalified necessarily, there are plenty of old tools that have been proven to be effective measurement tools and it's not necessary to go back and start over from

scratch with all this. There are some where there are problems that we find and in that case they should be addressed with appropriate qualification.

But we're not asking you to start over from scratch on everything. I think what we're trying to do is provide guidance that will allow us to, when an application comes in, it's actually in good shape. If the studies have been done appropriately we can interpret them, otherwise it's just, it slows down the process of drug development if we get an application where the studies were done with a tool that we can't interpret, and although we try to give as much guidance during development as possible, that doesn't always happen. Some companies come to us with their phase III studies completed and in other situations they just don't listen to us. So this is guidance, that's all it is and it's not intended to redo the entire panoply of tools that are already out there.

TURK: Thanks for the clarification. This morning, prior to the break what we tried to do is give you an industry perspective and I think two things became crystal clear to me, and hopefully to you, which is that the FDA is extremely thoughtful about the process, the steps that they go through, and they make tremendous use of expertise, including expertise of many people in the

audience. They read the literature, they go back to the publications that are coming out in relevant areas and based on their experience and that information, they come to some decisions about the processes they are going to use. And as Bob Rappaport just clarified, that this is a guidance; things to be thinking about when you come to them based on their experience, their knowledge that they've acquired from the experts in the area, including people in the room, their knowledge from publications. That's how they come to their decisions, they don't just make it up and toss dice and hope that maybe they'll come up with some kind of magic.

So one of the words that we sort of bantered about this morning was physical performance, as if we all know what it is, what is physical performance? So what we're going to be doing in the next couple of presentations before lunch and probably beyond, is to begin starting to understand what is this concept of physical performance, how do we think about it, what might be entailed. Some of it's sort of trickled up from some of the discussion, but I think we'll hear a lot more of it.

I'm delighted to have our first presentation by Dr. Dorcas Beaton who is the Director of Science and Scientists and Mobility program in at St. Michael's

Hospital, and also from the Institute of Work & Health in Toronto. Dr. Beaton, we luckily had her at one of our earlier impact meetings several years back in which you gave us a presentation on sort of a different perspective of things, and now we're going to sort of hear an update, and an update on what is physical function. And you realize no pressure on you, but Dorcas, this means you're going to define the entire rest of this meeting.

DORCAS BEATON: Thanks. So I'm happy to be here, I was actually happy to be on the concept side of it because I think some of the other speakers who are talking about the measurement will have a big challenge. So I'm here to talk to you about physical functioning from a conceptual point of view.

Okay, no financial disclosures, but I do want to acknowledge and disclose that I am an occupational therapist and that will probably come through in some of the ways I talk about people and their functioning. And I'm a developer of the not for profit probably at loss DASH outcome measure (laughter).

So physical functioning, it's importance in maintaining health and wellbeing has been well established way back into the 1980s if not before. Literature that we're starting to see now is that people, we have a

shifting cohort, that people are aging and living longer in a healthier state. So we're sometimes seeing compression of physical limitations into the very old.

The article that I sent out, the Tomey article, as well as many other articles of associated physical functioning to decreased quality of life, increased risk of disability with falls and fractures or depression, and increased health care costs. It's a very prevalent issue, and I went on the CDC website just this week, I found that 15 percent of American adults over the age of 18, that's 35.2 million people, have a physical limitation, and this prevalence becomes much higher, 43 percent, with many more number of physical limitations as you get into the older years.

Though when measuring physical functioning there is always a twist, that many people might express a limitation, they might have difficulties with mobility or in executing a certain movement pattern or a task, but then describe themselves as actually functioning quite well.

Let me give you an example of some work I'm doing with a graduate student of ours, Ellie Pinsker back in Toronto, and she says that when working with people with ankle arthritis, so often we'll put them in the labs and they can't ascend stairs in a lab because of pain in the

dorsiflexion. But then when we ask them in the community, so how are you managing doing a flight of stairs on the SF-36/PF-10 for example, no limitations, it's fine. So we say well how are you doing it, well they're walking up the stairs backwards because that doesn't require as much dorsiflexion so they avoid their pain. So they figured out different techniques to manage that.

The other thing they describe is the sense of having to be vigilant about their terrain because there might be a tripping hazard. And when they're walking, if they trip that's extremely painful, if their toe gets caught on something.

So we talk to them about social activities and social functioning, and here's where we see another twist, because if it's an evening backyard barbecue they won't attend because once it becomes dark they can't use their vision to compensate for their physical limitation and for their ankle arthritis. So you can see all these different strategies that people are using that make them say, you know, I'm doing okay.

So phenomenon of physical function, and trying to think about it as a concept is probably a really good place to start, as we saw with the FDA speakers earlier, before we actually start thinking about measures of it. So when I

say physical functioning, just in your mind think about what comes to mind. So, of course, I've been doing this for the last couple of weeks at every meeting that I've attended and asking people and very quickly people would nominate things with this gleeful look in their eye, I know what it is, I'll tell you exactly what it is. But what I found was this great diversity of different things, from force play data to trying to find out about people's muscle functioning in the lab, to ADLs, meeting life demands, working productively in my workplace. So a lot of things came up. And I think what struck me is what Gary Donaldson said once, is that sometimes getting numbers is easy, but measuring is hard.

So my objective today is to try to work you through some different concepts of physical functioning and to describe three challenges or three things that I think we need to pay particular attention to as we move into measuring physical functioning: The breadth of a concept; the role of the context of functioning; and then the ability to capture what's important; and then to outline briefly some considerations for going forward.

I've given a number of key articles. The first one is the one that was in your package, I'm happy to share this list and it will be up on the website of just some

other articles that I really did lean on for preparing this presentation.

So the concept of physical functioning, itself, is not new to IMMPACT, and you heard mentioned earlier that in the early IMMPACT meetings in 2003 and 2005, it was nominated as a very important part of measuring outcomes in chronic pain patients. And it sometimes might be captured in measures that talk about interference of pain on my roles or activities of daily life, like the Brief Pain Inventory, the Multidimensional Pain Index, or that that's captured by more direct or more generic health related quality of life. The domains of physical functioning and daily activities, and of course the one we've seen already is the SF-36 PF10 as an example of that.

When I went onto the American Chronic Pain Association website I was very pleased to see that recently they conducted a survey that endorsed that physical functioning is one of the key things for people with chronic pain that's important to them, along with enjoyment of their life and things like sleep and fatigue.

So how might we define it. I found some of the most useful definitions on the groups that are really trying to operationalize this, and this comes from the PROMIS website, one's ability to carry out activities that

require physical actions such as self care, ADLs, to more complex activities that require combination of skills often within a social context. And that's when we get into work, sport, teamwork.

So physical functioning is also going to include the full spectrum, from people with very severe limitations to our exceptional and elite athletes functioning at a very high level. So we've got breadth and depth that we have to be able to cover.

Others describe it more at physical functioning testing, that we're actually going to be testing the abilities or capabilities to do an activity. The ability to move through the movements that are required for a specific activity, or impairments tending to be at a more integrated level. Dynamic balance, power testing. In this article by Reiman and Manske they described actually 10 or more levels that you could go to in measuring physical functioning at this type of ability approach.

Now I know that many of us have now become quite aware and are probably living, eating and breathing the ICF, and we know that through this ICF model that was endorsed by the World Health Organization in 2001, that physical functioning can manifest itself at three different levels. It can manifest itself at the level of impairment,

which is maybe where we see some of the movement and things, or in activity, the acts and the tasks that we engage in in our daily life. Or participation, one becomes socially involved in different roles.

And we're getting pretty good at recognizing that when we look at these levels, we're actually looking through the lens of the disease, itself, or the disorder that's imparting its limitations on a person. So we get, we're quite used to saying that and seeing that. But there's these two little boxes at the bottom, the person and the environment, and in my experience I don't see us embracing that quite as much as being a defining lens for a level of impairment, activity limitation or participation.

So what I'll be focusing on a little bit more today is how do we bring that environment to the contextual factors up into our measurement of physical functioning. Because they are inextricably linked to the indicators of physical functioning that we're using. The PF10 is inextricably linked to where the person is in walking up those stairs, the design of the stairs, the availability of a handrail.

So for that reason I was very attracted to the article that I did include in your handout because I think it has an answer that we might want to banter about with.

Because what they've done is taken the concept of physical functioning using the ICF and then flipped it and rolled it right into all the environmental factors.

So there's three things on here that I want to sort of highlight for you. The first is that there's two rulers that we can measure people's physical functioning at a capability, what are you capable of doing, that might be some of your motions, your grip strength, some of the activities that you could do. And when we measure it out in the PF10 we're actually measuring it out here at performance. So how are you doing, do you have any problems doing that?

This diagram also recognizes that all of that, the performance is going to be the capability imbedded in the environment, you cannot see performance without environment, and that they define the environment as sometimes our usual environment, the things we're used to, so I can make my way around my house in the pitch black because I'm used to that but put me in a new environment and I'll walk into a wall because I'm not used to the new environment. And maybe the broader environment of things like the environment and the neighborhood effects, that there's curb cutouts, or safety features or maybe it's a weather issue for different environments.

That the environment has a different level. If you have very high capability, the environmental press, the fact, the amount that the environment is going to affect your capability and later your functioning is really limited because you are actually functioning quite well. At the very, very bottom, perhaps limited too, because you are bedridden so the environmental adjustments won't occur, but the other thing this does is brings another lens into our thinking about contextual factors.

People in the mid range of capability can go across and they can enlist adaptations, coping strategies, or they may decide to drop or avoid certain discretionary activities because they just simply can't do them anymore. And then that might, in fact, raise their level of performance on one of our questionnaires because they figured out a technique to adapt and allow their performance to be better. There's our ankle patient who says they're walking up the stairs backwards, so their performance is great, their capability might be limited, but they figured out a strategy.

Now in adopting that there still are some challenges that I want to go over, and the first is just the breadth of the topic they're actually talking about in measuring physical functioning. When I'm teaching

measurement, sometimes it tell people that a given outcome or given thing that you want to measure is like one window in a house, that it's giving you a certain view and it can only give you the view that's defined by the breadth and the height of that window and what side of the house the window is on.

So when we're thinking about measuring, for example, pain intensity, which I know we're not allowed to talk about now (laughter), but just as an example, we could measure pain intensity versus pain interference with my daily activities. And those are two different windows.

Now when we think about physical functioning I sort of think it's probably like the Versailles of concepts because there's so many different windows that we need to think about. And it may serve us well to start separating the windows out a little bit and make sure we have good measures of each window. So I mentioned to you what I encountered when I did my little informal survey of what physical functioning was. These might be different windows.

The components of physical functioning we might see: Strength, active range, reach, balance, movement patterns, proprioception, cognition, being able to do exactly the functioning to engage somebody in an activity in the community. And I mentioned vision as an example of

in the dark the ankle patient can't go out into an evening barbecue because they can no longer be vigilant about their environment. So it's the person's skill set that is enabling physical functioning, and that's one level.

Then we might move into some of the performance based tests. The timed tests, we've talked about the 6 minute walk test, another one is the timed up and go. And standardized tasks that have specific instructions or rules that we're trying to measure.

But even within that, there's huge variability. Here's two examples, a 3 meter walk test. The first one, timed up and go, you get up from the chair, walk 3 meters, return, sit down in the chair. If it takes, in this one study, over 15 seconds, it's considered to be a risk for falls in community dwelling elderly.

Take the second example, the zigzag run test where it's the same 3 meter test but now you have to go down 3 meters across 4.8 meters, around some pylons, back and if you do that in over 6.86 seconds you're above the normal limits. So you can see now we've got two tests, one much more complicated, both of them a 3 meter walk test in a way, with very different standards and trying to capture very different subpopulations in terms of physical functioning.

And there's other tests, too. This is the walking index, I don't expect you to read it, but it's the degree of assistance that somebody needs to ambulate 10 meters and it was designed for persons with spinal cord injury, zero meaning they're unable to do it, 4 means they can ambulate 10 meters using parallel bars, braces and physical assistance of somebody else, and 20 being that that is an independent support for 10 meter walking. So now we have the walking distance defined and the amount of support defines a level of physical functioning.

We can move into patient reported outcomes. So traditional paper and pencil tests, and we've talked about the physical functioning 10 item from the SF-36. or pain related, pain attributed physical functioning limitations. Or more regional scores that might have pain parts to them, like the lower extremity functional status, or something like the DASH. And what we're seeing with this is that paper and pencil is very quickly being replaced by computer based testing, apps, iPhone applications, and computer adaptive testing, and I'm very happy to know that there is somebody from PROMIS here so I don't have to tell you a lot about it. But the PROMIS initiative is one example of trying to provide widespread computer adaptive testing and infrastructure to support outcome measurement in clinical

practice and clinical research. And through PROMIS a number of items are ordered based on your probability of answering the next question right or the previous question wrong, and then it's able to zero in with a much fewer number of items onto a level of physical functioning.

There's been a lot of work, physical functioning is one of the domains that's measured in PROMIS and it was one of the core domains right from the very beginning. And there's a lot of work out including one out just this month in *Journal of Clinical Epidemiology* going through some of the psychometric properties of the PF items, there's 124 PF items in there that can cover a wide range because you don't have to ask all the items, you can zero in on those that are important and relevant for the person. And it's showing good reliability across a much broader spectrum of functioning with lower floor and ceiling effects. So things like this make it potentially more feasible for people to do a wide spectrum of physical functioning assessment.

But I would be remiss if I didn't talk about some of what's called now the instrumented outcome measures, the computer based reading of physical functioning and now we're having smart detectors built into clothing for heart rate, for movement, for upper extremity patterns. Sensory monitors, tracking devices in shoes, or even the simple

life line accelerometers that are placed on elderly people to track falls and then automatically elicit support if needed.

So I don't know how many people have a Nike field band on right now, or maybe they've got something tucked in their shoe to measure their steps, or a Fitbit in their pocket. So these are all things that make measurement of physical activity possible using technology.

The second challenge I'd like to talk about is the challenge of context. And as I mentioned earlier, everything we measure when we look at physical function is contextualized, just as it's driven by the disease or the disorder, itself.

So think about this, asking a question like how much difficulty did you have taking one small step, and maybe the response from two different people was well I needed some physical support. But context really matters, whether you're just building the capacity or you're one of the most physically fit people on the planet or off the planet, with exceptional cognitive and sensory ability, the absolute elite person is still feeling the press of the environment when all of a sudden they're having to fight because there's the absence of gravity whilst other people are fighting against gravity as they learn how to walk.

Now I come from Toronto where the press of the environment was really, really obvious this winter and everybody and anybody was grappling for whether it be a four-wheel drive, clamp-ons for your shoes, or walking sticks to make it to work on a given day this winter. And I think that a winter like this really tries to impress upon us just how vulnerable we are to the environment. The 60 percent increase in the fractures that occurred this winter, a lot of them shoulder and wrist fractures because some of the frail elderly did not go out in this weather.

So the presence of contextual factors are what define the situation in which our abilities are being tested or quantified. Whether you think it's there or not, the person who is responding to your questionnaire is answering from a context. It's part of every single estimation, whether it be ADL, work, recreation, it's a habitual situation, not always named or recognized but always present.

Potentially less so when we think about testing of grip, strength, power, aside from the immediate equipment or laboratory settings. And McNeil et al in 2006 described things like the quality of the neighborhood, the availability of facilities in a neighborhood impacting physical activity. Now when you buy a house in Toronto,

you get a walkability index so you know how walkable your neighborhood is, or whether you're going to have to be driving your car everywhere.

But it also defines things that are called upon as a resource or challenge to improve functioning. A walker, adaptations, technique, supports, or dropping an activity.

The ICF says that there's different chapters or domains within environmental contextual factors, like products and technology, the natural environment like our weather, supports and relationships, both practical and emotional support for different activities, attitudes, and services, systems and policies. So these are all things that are in place that affect somebody's physical functioning from an environmental one.

But let's think about the person as a contextual factor. And I call upon the work of Monique Gignac who is one of our people who is going to be speaking today who has done a lot of work on adaptations and coping in people with arthritis. That they'll do things like they'll adjust the time that they're doing, maybe inside versus outside work, they might use a vacation day instead of using a sick day to avoid a work disability absence. It's embodied in their willingness to seek help from other people. To modify

activities in order to conserve energy or make things easier. To anticipate and then avoid difficulties by organizing, pacing, or taking breaks. And also to drop activities.

In one survey of 248 people with arthritis, only 3 people reported not having some sort of adaptation technique thing that they did to improve functioning. And probably a lot of us do that, too, when we put on our glasses, or when we use a certain pair of shoes for running.

So we have to be aware that some instruments penalize for the use of these assistive devices and that will be one important thing that we need to think about as a group, is that good because it's allowing the person to function better in the environment, or is it something that we want to penalize against.

Ability is in the eye of the beholder, and I learned this very quickly when I was doing some work in hand therapy with people with Charcot-Marie-Tooth where they gradually lose intrinsic functioning in their hand. And we would say to this person are you having any difficulty picking up small objects; nope, everything is working fine. They had other concerns about the atrophy and the look and appearance of their hands, but picking up

small objects like this penny as not difficult.

So there are measures that are designed to capture that. They ask your level of independence without assistive devices and then ask a second question about your access to resources to overcome that limitation and is that satisfactory in your life. Is it accessible to you on an ongoing basis. So that might be something to think about.

The third challenge is the challenge of capturing what's important. Noble, in the 2013 article, the one I point you to there, where they talked about the universe of activities, the universe of every possible activity that we could have in a given measure. And that might include the larger blue box here, the things that I'm capable of doing. There's lots that I'm capable of doing. In the smaller red box, the things that I actually did today. Maybe some of that's the things that I had to do, I had to get out of bed, I had to get dressed. And maybe others are more optional things, some optional discretionary activities I might like to do during the day.

So those encompass the things that we typically have in some of our measures. But beyond that, we should be pushed to think about the things that I used to be able to do before I had this condition but I had to drop them. The things that I really would like to reassume because that

might be setting the expectations for our patients. And maybe the green box, the things that I aspire to do, once I get over this issue I really want to be able to do this or that. And there might be another array of activities that they're hoping to do. And that's only unless we capture that full spectrum of activities, the holy grail as Noble says, as measurement of physical functioning, that we can really be sure that we've captured the whole experience of physical functioning, the floor, the ceiling, the expectations of physical functioning of our patients or the people that we're trying to measure physical functioning in.

But that also comes at a cost of practicality because often if we have that we need a huge array of items or a way to access those items. And PROs are often moved into short forms by dropping out the high end items which will be the vitality, the things I want to do, the high end performance items. Or by selecting to space items along a ruler so that we do have the length but maybe we don't have the precision and the detail. So maybe the item that I really value is dropped out and it's just not there.

Now items banks and computer adaptive testing allows for this wider array of item banks but that still might be focused on the items that are usually done during

our day. So can we use those items banks to pick up the things that I aspire to do or the things that I had to drop doing.

Patient specific approaches are ways where we ask the patient to actually elicit their own items for a given scale and the items are nominated, and they tend to reflect current activities similar to what our other measures do, but they also reflect those items that they're aspirational items, or the items they had to give up. And they would allow full spectrum but we run into measurement issues like how do we compare people who have a different set of items in their tool.

So in summary, measuring the concept of physical function is important. It's driving a lot of things in our health care system but it's also important to the people with chronic pain themselves. When we do that, there's certain things we need to think about, the breadth, the variety of windows we might need to have to be able to capture this concept of physical functioning. We have to always be aware that the context is there, it's there to name it and measure it, might be more helpful than to just ignore it and then wonder why we have these differences in scores or similarities in scores when people are obviously quite different. It defines the situation in which the

person is exerting their effort, their capabilities, but it's also defining the factors that can be pulled in to help somebody overcome a barrier.

And finally, we need to consider the meaning. That many measures might miss what is really important to people and what's the activities that they conceive as being important. And can we integrate in some of these extra important items, the aspirations, discretionary activities, we'll hear a talk about recreational and social activities today to get a fuller picture of physical functioning.

So I leave you with this model to think about because of it's ability to think this is the actual figure from the article that's in your handout, because of it's ability to pull out and separate this capability, the actual performance that we measure and then the environmental impact and contextual factors.

Thank you.

TURK: We just decided that we're going to extend the meeting for an extra four or six days to cover some of the topics and challenges that Dorcas gave us because this obviously illustrates the complexity and the areas that we need to be thinking about considering. Unfortunately, we didn't plan it what way, but it doesn't

mean that we wouldn't consider there should be a follow-up subsequent meeting on this same topic, either an IMMPACT or OMERACT, or however some other variation that might be. But I think that was an outstanding breadth scope, gave you a sense of what's involved and why it's important, and why we need to consider these things.

I think Dorcas mentioned patient reported outcomes, she mentioned observer outcomes, performance based outcomes, and I think those are the kinds of ways that we could think about how do you go about assessing performance. Each one of those is not right or wrong, it provides different information and as I think you'll be hearing over the next several presentations where we go into more in depth in these, each within itself is a complex area with its own challenges and questions. And then by the end of tomorrow, we're going to sit back and try to see how do we pull this together. So you are not off the hook, we're not going to have four more days, so we're really shooting for tomorrow afternoon.

The next presentation is going to be specifically to focus in much more depth on patient reported outcome measures. Dr. Ann Taylor is going to be presenting. She's a reader in pain education and research in Cardiff University, someplace on the other side of the pond, and

she is going to really take us into the weeds, if you will, about some of these patient reported outcomes. It's also important to acknowledge that Ann has been given the task of being one of the two rapporteurs so she is taking notes and taking information that is going to help guide us, and we will save questions for both Ann and for Dorcas, as well as one of the other presentations until after the lunch break, and we'll have all three of those speakers come up and we can talk more in depth about these things.

So Ann, you're going to help us deal with them or ask them to be challenged.

Patient-Reported Outcome Measures of Physical Function

ANN TAYLOR: Hi, thank you very much for asking me to speak today, and I started looking at the subject and I thought this is really easy, I know what I'm talking about, this is going to be -- and then I started looking at it. And the more I looked at it, the more complex it became, the harder it was to define if we were going to do a systematic review round physical function, physical activity outcome measures, what is it we were actually going to look at. And I want to explore some of this with you today and given that I'm the fellow and you're the experts I thought you could help me dissect where we need to go with some of this.

So as we all know, people living with long-term conditions can experience physical activity limitations or suffer from increased symptoms during activity. Now the reason I've put a long-term condition there is now within the United Kingdom we actually have government endorsement that chronic pain is a long-term condition in its own right. So as a result of that the people with chronic pain, people living with chronic pain can actually have the same resources as those with diabetes, with heart disease and chronic chest diseases, things that they have never had in the past because it's very much been a Cinderella type

of issue with a lot of people with chronic pain actually not being believed. So that's been a great leap forward.

And I think part of the physical function patient reported outcome measures issues is we need to be able to compare how do people living with pain compare to people living with arthritis in terms of outcomes from drug studies or from any kind of study. And Spangler, et al, showed that patients with chronic pain actually had a worse quality of life than even people with cancer and I think that, we need to raise the profile here.

Physical activity is important in preventing and managing many long-term conditions, especially chronic nonmalignant pain and especially things like fibromyalgia. Physical activity outcome measures are useful in clinical trials because it enables researchers to effectively evaluate the impact of treatment options; what it actually means in terms of a reduction of pain. And I think there are issues that having listened to this morning and actually looking at the literature, is just because the pain is reduced doesn't mean function is going to improve. And I think it depends on the individual's ability to catastrophize their fear, their job status, et cetera, et cetera, about whether they actually will, their function will improve as a results of their pain being reduced.

There are a myriad of objective and subjective outcome tools, as we've already heard, and the aim of this presentation is to examine patient reported outcome measures that can be used in clinical analgesic trials in patients with chronic nonmalignant pain, chronic musculoskeletal pain. So I am going to give a brief overview of the domains covered and I split the outcome measures into various domains just for ease of presentation, but they might not be right. The overview of the differences and similarities in the content and methods of outcome measures in each domain, a brief discussion of the strengths and weaknesses and some of the key issues.

As already mentioned, IMMPACT actually has done a lot of work around pain and physical function, and in a patient or people who are living with pain focus group, they found that physical limitations and physical functions were very much altered because of people's pain with the majority expressing a real problem with how they functioned physically, although some did manage to climb stairs to get to their office, or stand for lengthy periods of time as a result of their jobs, things like bending and lifting still became problematic.

As we've already mentioned, there's the OMERACT filter that we need to bear in mind with truth,

discrimination, and feasibility, and the OMERACT 2.0 core outcomes set. So those are just as background context.

Now I'm going to use physical activity and physical functioning interchangeably because I'm as confused I think as everybody else about what it is this systematic view may look like in the future. But just for you to know that I do know the differences, I've put the definitions here. So physical activity can be considered as any bodily movement produced by contraction of muscle that increases energy expenditure above a basal level, where physical function has been defined as the ability to carry out various activities that require physical capability. And these range from self-care, which is the basic activities of daily living, to more vigorous activities that require increasing degrees of mobility, strength or endurance.

And just for completion sake, there's a definition of patient reported outcome measures. I've included any report of the status of the patient's health condition that comes directly from the patient without interpretation of the patient's response by a clinician or anyone else. It can be measured in absolute terms or as a change from a previous measure.

So in order to inform this talk I did a search

and for those cosmonauts out in the audience, you'll see that it's very, very woefully poor, but it was just to kind of get to some idea of what is actually out there in terms of physical functioning measures.

The assessment of a physical activity as an outcome measure does provide a unique perspective in chronic disease research, not only for observational studies but also for drug and non-drug studies. Furthermore, evidence from clinical trials regarding physical activity as a patient reported outcome could inform patients about treatment. Again, we've got massive drive in the UK to actually start using shared decision making tools and unfortunately America has kidnapped one of our famous shared decision making tool developers, Professor Glyn Elwyn. But he's led a whole process of getting two (indiscernible) of A4 tools that you can give to patients so they can make a decision about their treatments based on evidence base. So physical functioning being so important to patients will actually, if we use the right tools to et the right outcomes, we can actually share this with the patients in the hope of them making good decisions about their care.

It is also investigators who are interested in measuring physical activity do face the challenge of not

only choosing an instrument that serves their study aim, but also one that's been carefully developed and validated, and the FDA shared some of those issues with us this morning. These instruments should have strong psychometric properties such as stability over time, and also the capacity to detect even small effects. And in addition, investigators need to be certain that the instruments reflect the dimensions of physical activity relative to the patients and I'll talk about this in a minute.

It is currently unclear whether available instruments to measure physical activity in patients with rheumatological and arthritic conditions fulfill these requirements.

So one of the first domains that I looked at and there might be measures missing here and I apologize for that, but just to say, you know, once we develop this, once we get feedback from you guys today, and from OMERACT 2014, then obviously search criteria will be far more robust. But these pain related physical outcome measures have been described in the literature. And as I've already alluded to, pain intensity and physical functioning don't go hand in hand. And the IMMPACT recommendation suggested has already been described to use disease specific tools where possible and also in conjunction with more generic

measures, and recommended generic SF-36 and also more specific the Brief Pain Inventory and MPI.

They identified really good examples of disease specific physical function tools such as WOMAC and the Roland and Morrison Back Pain Disabilities Scales.

Next we have general physical outcome measures, again, you can see there's a whole host of these available as well. And while physical function outcome measures can be useful in an attempt to document the range of disability in the general population or patient group, they may be unresponsive to actual disease specific changes which is of concern.

As I said, the SF-36 is the most commonly used generic measure of health related quality of life and it has been used in studies of diverse medical and psychiatric conditions, and published in numerous studies. However, it lacks sensitivity and specificity as we've already discussed. And while this tool does cast a wide net, there are new developments in quality of life -- sorry, health related quality of life tools, and they may offer improvements over SF-36 and may actually eventually replace them. So we have this generic general physical outcome measures.

We have activity of daily living outcome

measures, as well, and again, these may merge with the previous category, but again, it was just trying to get some sense to present on the screen.

So this is an activity of daily living, as you are well aware, it's an umbrella relating to self care comprising activities of task that people undertake routinely in their lives and be divided into basic ADLs and instrumental ADLs. So basic ADLs are usually restricted to use of functional mobility and instrumental ADLs concerning the abilities of individuals to cope with environment, et cetera --

Typically, treatment goals include achieving maximal increases in function or participation in everyday life for the patient or client, and functional assessment is the method used to document these outcomes with activities of daily living skills being the most frequently used tools.

We have a whole range of disease specific activity tools, as well, but unfortunately, not all of the diseases that we are now managing in clinical practice has a functional physical activity or functional outcome scale associated with it. These, however, are probably the most interest to patients and their treating clinicians compared to the generic measures because you are actually drilling

down to what does the disease actually do for your physical function or your physical activity levels. But you can't then compare these disease physical function measures with other diseases, they probably won't relate. And there's a number of site specific physical activity measures as well which have the same issues as the disease based ones.

So there are certain burdens on individuals. IMMPACT recommends there should be a disease activity measure combined with maybe a more generic measure in order to compare across populations. But this actually, and they do actually say that it is useful to use a number of measures as long as it's not too burdensome on the individual. And this is the problem of, you know, to get your utopia of physical outcome data, how many tools are you going to have to use in order to get that kind of information that you want.

Also, statistically, how are you going to manage controlling for the multiple datasets that you are going to get because 1 in 25 is going to give you a positive reading. And there's also the issues of pragmatic versus, you know, theoretical approaches that we want to be performing individualized care so we want data on how to do that individualized care. Grant awarding bodies wants us to incorporate that in the grants we're going for, but when it

comes to publish the evidence or actually seeking further work, it may be difficult to actually be as pragmatic as you'd want to be.

So there's a number of systematic reviews that have already been done on patient reported physical function/activities. Most of these have been done using COSMIN. There's a number of hip and knee ones, there's ones associated with a younger age group. There's a generic one on physical activity. There's one on neck pain and disability, one on hand pain, and one on hip arthroscopies. So the systematic reviews have already been done looking at some of these functional tools.

So some of the key points that I've pulled out from looking at some of these physical function/physical activity tools. No outcome measured or key domains, and that is an issue. Some questionnaires focused on physical activity alone, others looked at physical functioning and some included multiple domains which physical activity or function was one or was a subscale.

Questionnaires tended to be developed for patients with long-term conditions and many focused on the older adults. So again, so you've got floor and ceiling issue here. The format of the questionnaires varied considerably, but most were unidirectional, self

administered and scored by calculating some of the domain or total scores. Many of the questions were developed for a range of populations and limitations experienced by some groups were not applicable to others.

There appears to be no consensus on what should be included in terms of content and format in patient reported outcome measures for physical activity, especially associated with painful conditions. Previous reviews have found variation in the number of recall periods, used inconsistencies in the development and validation methods of questionnaires, and conceptual frameworks for physical activities are scarce, and this may explain the lack of consensus.

The other issues that I was considering this morning listening to the talks is well, what are we going to combine physical function methods with. As I said, it doesn't actually equate that because you reduce pain you're going to increase physical functioning because patients still might be fearful, they still might catastrophize about that pain. So should we actually be recommending that if we use a physical function outcome measure or set of these measures, they should be accompanied by aspect that measure fear and catastrophizing.

We've also got the issue of in the UK certainly,

our average reading age is 12 and it's going to be terribly burdensome giving patients a whole range of questionnaires that maybe they don't read terribly well. Also there's the evidence to suggest that in a lot of these painful groups they are of low educational level and low socioeconomic classes, as well, and that actually drives the burden that all of these tools are going to have when you are trying to actually find out how pain is impacting on physical function and how clinical trial treatments actually improve that function.

There are various checklists that are around for assessing patient reported outcomes. This is one by Guyatt, et al, in 1997, and certainly COSMIN addresses all of the issues here raised, what were they measuring, what are the admissions, what are the measurement strategies, issues with ability to measure change, validity and it's impact in clinical practice.

We've already heard the FDA guidance and I think this guidance actually put patients very much in the center of research. It's emphasis is placed on defining the roles that a PROM endpoint is intended to play in clinical trials, so the chosen outcome measures are matched appropriately to what you want to achieve. Secondly, the fundamental role of the PROM content validity is expanded

and clarified in this document. The FDA acknowledges that different approaches may be needed to insure the existing modified or newly created PROMs are appropriate to support product labeling.

The guidance also states that sponsors need to supply evidence that a PROM instrument adequately measures what it's claimed to measure. And obviously, also it needs to have a good developmental history to go along with this.

The guidance, as I said, put patients at the center of the treatment decisions, providing these measures are selected and interpreted appropriately, and related precisely to a priori hypothesis regarding treatment outcome. Moreover, this guidance ends the unscientific practice of including any vaguely relevant PROM measure in a clinical trial at the eleventh hour. Instead, a clear strategy for the inclusion of the PROM in a clinical trial program will be required just as further endpoints.

And while the FDA explicitly states that this is not for work outside clinical medicinal products, I think the basic principles are sound for any kind of research, know what you want to measure, why you want to measure it, and be sure that the PROM instrument you choose measures what it is intended to measure. And I think those are three of the things that I have problems with in actually trying

to decide what the systematic review should look like in terms of physical functioning matches.

So I thought I would ask you some questions, and I don't know whether we can take those now or whether we discuss them after lunch. What are we going to review in terms of patient reported outcomes? Are we going to look at physical activity, physical function, physical fitness, disability, et cetera, et cetera, all or some of those? And what domains are we going to examine? Are we going to narrow the focus and follow IMMPACT suggestion of having generic outcome measures and disease specific outcome measures, or are we going to look at just pain outcome, physical function outcome measures, and can we actually justify the narrowing of what we want to do? I mean I have no problem doing all the work, but I think it has to be a meaningful review that people will pick up and use. And if we include all these outcome measures, then it's just going to be a massive tone that nobody is going to want to look at. We could actually narrow the focus by looking at a disease condition such as fibromyalgia, that was another thought that I had.

Should we change the domains and actually look at categories, so look at having a review of those that are evaluative, those that are discriminative, those that are

predictive, those that are planning, but again, those barriers will merge as well.

How can we include what has already been done using COSMIN, and other systematic review tools, and should we actually use the ICF codes to assess for relevance? I mean certainly a lot of the tools that I've reviewed actually pre-date ICFs so it is difficult to use that. And there has been some criticism that it's not as exhaustive as it could be for a comprehensive classification.

Thank you very much.

TURK: I think given that we're scheduled for lunch from 11:30 to 12:30 it is probably best to save questions for what is a rather lengthy discussion period after the next two talks. But Ann, what I would like to arrange with the AV people is if we can come back to this slide at least during part of the discussion because obviously a lot of these questions would be excellent for discussion during our discussion period. So unless someone has a strong disagreement, why don't we break for lunch now, reconvene at 12:30 when we're going to have two talks, and then have an hour to an hour and a half for discussion of all of these challenging issues. But thank you.

TAYLOR: The talk was a lot longer than this but I actually cut it short because I was repeating a lot of

what was already said this morning.

TURK: We might even try and get this slide printed out so that everybody has a copy in front of them and can sleep on it.

TAYLOR: Thank you.

TURK: So let’s do that, thank you.

(off the record - lunch break)

TURK: Thank you, all, hopefully you had a lovely lunch and you are pumped and ready to move on. We had a very stimulating discussion by Ann Taylor in the discussion presentation and we've put out by your seats, you should have by your seat one of the slides that Ann had gotten together, some questions, and although we may or may not get to these today, they're definitely going to be something we're going to more be talking about tomorrow when we start pulling things together. So please hold onto that and we can use it to help guide some discussions that we're going to have.

Now up to this point, so we've talked about the industry, we've talked about the FDA perspective on outcomes and we suggested that this is broader than just for drugs, this is for any type of treatment. We then has Dorcas Beaton give us a tremendous presentation pointing out to why we've changed your checkout time on your rooms you can't leave until we give you permission which is until we resolve all the questions on this sheet. And not just for PROs, patient reported outcomes, but also for performance based outcomes, which you are going to hear about next, and also about observational methods which you are also going to hear about, and even some specific methods that are out there.

So you can't leave until we've got all those things resolved so plan on spending the summer in Washington, DC, I heard it's beautiful in Washington in the summer, people want to come here, it's a tremendous site to attend.

Okay, so what we're going to do now is switch from the patient reported outcomes to now moving at some more quasi objective or objective, depending on how we want to look at them, some laboratory based methods, some observational based methods, some clinician based methods, to try to learn, again, more about function, but from a different perspective. And remember that tomorrow we're going to try and pull all these things together as we move forward.

So I'm delighted to have our next speaker who is an OMERACT fellow, which I hadn't heard that term before, I'm delighted to have Dr. Kristine Phillips here. Dr. Phillips is from the University of Michigan and she is going to give us a perspective on another way of thinking about, of not thinking, another way of evaluating functioning by focusing more on more objective type measures. Dr. Phillips.

Clinician, Observer, Laboratory, and Other Outcome Measures of Physical Function

PHILLIPS: Thank you so much for asking me to talk today. I'd like to begin by just reiterating the definition that Dr. Beaton laid out for us, that with performance based tests and individual needs to execute one or more specific activities that are then evaluated in a standardized manner. And these are usually measured as time to complete the activity, but they can also be reported by an observer.

So why do we want to measure performance? It's important to remember that self report can be subject to under or overestimation in musculoskeletal disease, and this has been shown in a number of different musculoskeletal diseases.

There are really only moderate relations between both measures, between patient reported outcomes and performance based measures, particularly in the elderly. Time based performance and self reported measures likely assess different aspects of the physical function domain. Patient reported outcomes may assess, for example, the effort that a patient is perceiving that they're putting into the function.

There's evidence for discrepancies between

perceptions of the individual using the self reported measure, and their true ability. They may underestimate the performance that they're doing or they may overestimate it. And this can be due to a variety of things, including personality traits, language barriers, and other comorbidities like depression.

There has been a lot of discordance that has been observed between physical functions measured between self reported questionnaires and patients who have ankylosing spondylitis, rheumatoid arthritis, or fibromyalgia.

So the goals of this talk are to describe these performance based physical function assessments as outcome measures for a chronic musculoskeletal pain randomized control trial. I'm going to give you an overview of the different measures, I'm going to focus on global measures of physical function and touch on localized or segmental measures of physical function, and some of the complexities that are involved. And then we'll talk a little bit about the measurement properties of each of the measures.

This is a conceptual framework that describes the disability that people feel when they experience musculoskeletal pain. And this talk is going to focus primarily on the functional limitations and not on activity restrictions. Those topics such as actigraphy, difficulty

with doing ADLs, social participation, work participation, will be covered in other talks later today.

So when we assess the impact of musculoskeletal pain on functional limitations and we think about the core areas of measurement that we want to evaluate, it's important to think about the domains, the sub-domains of physical function that we want to look at.

This is a standard slide that I think Dr. Mease went over in the beginning, showing how we go about developing a preliminary core outcome measurement set. If we have instruments within the core domains that we want to look at, we then apply the OMERACT filter looking at is this a truthful measure, is it discriminative, and is it feasible.

It's important to remember that within the core domain of physical function there are other individual contextual factors and these were touched on by Dr. Beaton in her talk. These are things like effort, how much effort a patient is putting into their function, their perception of exertion, and then other comorbidities, when we're talking about 6 minute hall walk test, for example, if there's a decline in the function measure this may be due to pulmonary arterial hypertension or interstitial lung disease, something that we're not as interested in

measuring.

So why do we use rheumatologic studies? There are a lot of rheumatologic studies that use performance based physical function outcome measures to form a core set of endpoints for clinical trials in a variety of diseases, including osteoarthritis, rheumatoid arthritis, fibromyalgia. There have been more localized function measures in low back pain, and then a lot of segmental or joint specific musculoskeletal pain outcomes that have been developed for shoulder pain, knee pain, that sort of thing.

So just as an example of one of the best studies that's' been done in this area, this is a recently published osteoarthritis systematic review looking at measurement properties of performance based measures in patients who have hip and knee osteoarthritis.

They did a standard literature review that was very rigorous in its exclusion criteria. They were able to whittle it down to approximately 24 studies, and then they applied a modified version of the COSMIN filter to assess the quality of each of the measurement properties that were used in the study. And the studies used multiple measurement properties. They looked at reliability of each measure, internal consistency, responsiveness, measurement error, validity, including content, structural, content,

cross-cultural and criterion validity.

And then they assigned a rating of the quality for the level of evidence for the measurement property, and compared each study. And just to go through these briefly, there was a strong level if the rating, you know, showed consistent findings in multiple studies or good methodologic quality, or one study of excellent methodologic quality.

And this is an example of a study that looked at a 50 foot fast paced hall walk test. Pretty average for the studies that they found, there was not really any measurement error, they did not test validity, they did not test responsiveness. They did test intra-rater and inter-rater reliability, but not test/retest reliability.

So are there performance based physical function measures that are used across multiple musculoskeletal diseases, and if so, how do they compare. So there are a couple, there's the timed up and go, or the tug test that is used a lot in osteoarthritis, and it's been used in fibromyalgia as well. It incorporates walking down a 3 meter hallway, turning and then returning to sit down. So it's an assessment of both walking and turning.

There is another variation that's used as well, which is the get up and go test, which is just walking for

20 meters with no return. And this illustrates how it's done with the observer measuring how long it takes the patient to walk the 3 meters and then turn around and sit down.

So going back to the OA systematic review results, there were six different sit to stand tests that they looked at and there was a lot of variability in the methods that were -- there was a lot of variability in the measurement methods. They varied with the number of times that people did the assessment, the total time, the height of the chair, and they incorporated sometimes walking and sometimes turning components but it wasn't consistent. And they also used a sit to stand test in three multi-activity measures.

Another rheumatoid arthritis study looked at the same test and this was an interesting study done for a randomized control trial of rehab in patients who were sent to either Norway or to the Mediterranean to undergo rehab. I thought this was a really interesting way to look at this (laughter).

MALE VOICE: Can I volunteer?

PHILLIPS: Exactly, I'd like to run this study. So patients showed an improvement in ACR20 which for those of you who don't speak rheumatology, that's a 20 percent

improvement in the tender and swollen joint count and a 20 percent improvement in 3 of the 5 ACR core measures. And those are patient and physician global assessment, what we've been talking about all morning, pain, disability, and then acute phase reactants.

So while 25 percent in the Mediterranean climate achieved ACR20, only 15 percent in the Norwegian climate achieved ACR20, but there was no difference seen in the TUG test between the two groups. So this illustrates one of the potential pitfalls of these performance measures. Over and over again they're really not always associated with disease activity, and this has been shown multiple times for RA and for ankylosing spondylitis.

The 6 minute walk test has been used a lot for cardiopulmonary disease as an outcome measure. There are even guidelines from the American Thoracic Society for how to perform it in a standardized way. And building on this, you know, prior success is very appealing, documentation should include the speed tested if the fasted speed is not used, so either preferred speed of the patient or the fastest speed reached.

Assistive devices can be used so you are not limited by that, but it needs to be kept consistent from test to test. And the length of the track that's used

should be taken into consideration when you're doing the calculations.

So again, going back to that osteoarthritis systematic review, there were two main types of walk tests, there were short distances of less than 100 meters, and long distances of greater than 100 meters. They had 9 different short distance walk tests with lots of different variations, variations in the set pace, variations in the distance walked, and variations in the functional measure. And then the number of incorporated turns in the walk test also varied quite a bit.

The short distance walk tests were included in six different multi-activity measures which don't always, they say that the multi-activity measures typically measure physical function, but sometimes they measure different domains of physical function and that is one of their limitations.

The 6 minute walk test was really the only -- the 6 minute hall walk test, as it's also known, was really the only long distance walk test and was investigated in four studies and included in two multi-activity measures.

So this 6 minute walk test has also been used to assess physical function in fibromyalgia. This was a study done in lean patients with fibromyalgia. They have also

used it in obese patients, but there were limitations on that study that because of the 6 minute walk test that was utilized. So this was a better study to look at and they were able to demonstrate that patients with fibromyalgia who were of equivalent BMI and equivalent weight to healthy controls walked shorter distances during the 6 minute hall walk test and had comparable levels -- had higher levels of pain, rather.

Then there are a number of different stair negotiation tests that can be used and these can vary in height from 4 stairs, as you see here, to 9 stairs up and down, or 12 stairs up and down. And when you look back at the osteoarthritis systematic review results, there were 7 different stair negotiation tests that they found showing that there were variations in the number of stairs. There were variations in how it was conducted, whether they were ascending only or descending or both, whether or not patients were allowed to use the handrail support, and then the leading limb step pattern which also can have an impact on the outcome. And then there were five or six of the multi-activity measures that had, depending on how you counted, the stair negotiation test included.

There are other physical function measures that are also used for osteoarthritis or for localized

musculoskeletal pain, such as low back pain. And this includes the one leg hop, which was shown in a separate systematic review to be the most sensitive test for patients who were young or middle aged and at risk of developing osteoarthritis, or the standing stork test which is what you see here that looks a little bit like a yoga position.

And then there's the chair stand test where a patient is sitting in a chair and then rises to a standing position five times, and they have 30 seconds to do this measure.

There are other measures of balance that are used as physical function measures. This is one that's been tested in fibromyalgia, the Fullerton Advanced Balance Scale. This is one of 10 measures total that are assessed by an objective observer for a total score of 10 points.

And then there are site specific physical function tests, such as the loaded forward reach test that's used for low back pain, shoulder range of motion test, knee range of motion test, knee, quadriceps strength, hand grip strength test which is used a lot for hand surgery outcomes, and then the single leg hop test that I alluded to earlier.

The multi-activity test of physical function are

numerous and there are only part of them listed here. They're really subject to differences in measurement in different domain measurements. So some of them may measure range of motion of the neck, some of them may measure range of motion or physical function of the lower extremities as one battery. So they're limited because you're really not measuring the same thing with all of those different indices.

There has also been comparison of specific physical functions and this has been done in ankylosing spondylitis where they looked at the Bath Ankylosing Spondylitis Functional Index, which is a questionnaire that corresponds to disease activity in ankylosing spondylitis patients. And they looked at performance based tests based on those questions which include how difficult do you find it to climb stairs, bending over, reaching up, putting on your socks, reclining and declining from a chair, getting up from the floor without assistance, turning your neck and looking over your shoulder, and then any physically demanding activity.

And similar to other multidimensional assessments, everything except test 7 corresponded to general overall physical function. And there was a moderate association seen between the questions that patients

reported and the actual measures that patients were able to demonstrate that they could do.

So I wanted to just also mention that with regard to physical function, when we're measuring patients there may be discrepancy between their perception and what they are able to actually do. And this is really due to adaptation, as Dr. Beaton illustrated in her patient with ankle pain. This is sort of the creation of a new normal. The performance is better, but the capability is still limited and it's because they over time have gradually had a decline in function but are no longer able to do what they typically could do before.

So patients may really underestimate the levels of physical function that they have due to disease duration. And it's important to remember that patients with longer disease duration may tolerate greater limitations in function because they're not perceiving that slow decline. Each year's a little worse in overall function and the decline is faster than they would have expected or than they perceived.

There are other considerations whether performance based physical function measures are truly more objective than other patient reported outcome measures. This is, you know, when you compare the amount of time that

it takes for someone to do a physical function measure compared to an objective, an observer rated measure, you're really measuring something that is objective like swimming in the Olympics to ice dancing, which is more subjective, and there is an observer who is rating that assessment and subject to bias potentially.

And then finally, another consideration of the physical function test that we have consider includes sensitivity to activity. There have been a number of recent studies in osteoarthritis that suggest that certain individuals with chronic pain conditions may have increased sensitivity to physical activity and this can impact their measures of pain. This highlights this activity related pain among individuals is important, not only for osteoarthritis patients but also for low back pain patients and fibromyalgia patients.

The sensitivity to activity was significantly correlated with the 6 minute walk performance test such that higher levels of sensitivity to activity were associated with reduced distance traveled. And that's illustrated in this slide from Michael Smith's work that was recently published looking at the measures of activity -- I'm sorry, the measures of pain with repeated 6 minute walk test duration.

So what would an ideal physical function performance measure look like? It would be related to patient symptoms, as we discussed earlier in the morning, it would measure only one domain, only one outcome. It would be sensitive to change, and accurate, free from error, and easy to perform ideally.

There have been a number of studies that have measured physical function in a so-called objective way, but many of them are limited because they include multiple assessments of different domains, and there is really limited reliability and validity data. There were only a few 6 minute walk tests or timed up and go tests for patients with knee or hip osteoarthritis that could even be used in the very well done systematic review on osteoarthritis.

It's really important that we take away from this that future studies should include measurement properties for performance tests and we have a challenge before us to develop more performance tests that can discriminate between the therapeutic and non-therapeutic tests on patients with chronic pain on physical function.

And these are some references. Thank you. I'm happy to take questions now or we can -- wait till later? Okay.

TURK: Well, Bob Dworkin and Phil Mease, we made a big mistake, because we're going to cover all of this in a day and a half meeting, perhaps we should have thought about one meeting for physical function, for patient reported outcomes, for observations, but the good news is we'll have other meetings, and, therefore, you'll all be able to come, attend, and participate, and address some of the kinds of questions that are being posed. And our speakers have really posed important questions really for us to be considering.

Now, we've heard several times about performance based measures and quasi objective or objective measures, and the word actigraphy has been mentioned for the aficionados who are familiar with that, they knew exactly what was going on. For some of us who are a little bit less familiar with it, we thought as an illustration, not that actigraphy is the only objective measure out there, but rather let's pull it out and look more intensively at actigraphy to understand what this type of measure, and Dorcas Beaton showed you a whole range of other kinds of measures, electronic measures that are out there, but just to use actigraphy as an illustration to the complexities, some of the issues, the importance of looking at that, and speak to some of the concerns that we heard raised about

actigraphy but I think could be generalized to any one of these more electronic approaches.

So I'm delighted to introduce my colleague from the University of Washington, Dr. Kushang Patel. Dr. Patel came to University of Washington and I was fortunate enough to recruit him from the National Institute of Aging where he was functioning heavily as an epidemiologist, but also was involved in clinical assessments of older populations, particularly physical functioning, and had made lots of use of a number of the tests, the timed up to go, the 6 minute walk, but also had spent a significant amount of time looking into actigraphy and using that in some large national studies.

So Dr. Patel is going to take up the issue of just pull out actigraphy as one type of performance based objective measure and what can we learn from it, what do we know about it, and how does it extrapolate then to the other kinds of devices that might be out there that we would also apply the same kind of issues. So Kushang, they are yours.

Actigraphy

KUSHANG PATEL: All right, well thank you to OMERACT and IMMPACT for the opportunity to present this afternoon. Are my slides available? Thank you. So I mean I'll be talking about actigraphy, as Dennis said, it's an objective assessment of physical activity which is, as we all know, is a very difficult measure, or difficult behavior to measure as it is unstable. It's prone to daily, weekly and seasonal variation.

So the objective of today's talk is to provide an overview of accelerometry or actigraphy as an objective measure of physical activity for use in analgesic clinical trials.

So what are accelerometers. They are small, lightweight, portable, noninvasive and nonintrusive devices that record motion in 1, 2, or 3 planes, so that's uniaxial which is typically measuring vertical displacement, biaxial typically measures vertical displacement and anterior/posterior displacement, and triaxial accelerometers also measure, additionally measure side to side or mediolateral movement. And these measures, the accelerometers measure the frequency, duration and intensity of physical activity. These devices are increasingly getting smaller and smaller. This is an

example of an accelerometer sensing unit, it's smaller than a quarter. They typically weigh less than 1 to 2 ounces, so 3 to 4 quarters, so these are quite small devices.

So I thought I'd start out being a little bit provocative by displaying some data from the National Health and Nutrition Examination Survey data, which is the Centers for Disease Control and Prevention's data source for monitoring a variety of behaviors and a variety of disease outcomes. So these are data from the 2005-2006 wave and in NHANES they've taken great care to try to come up with good measures of physical activity that capture not only leisure time physical activity but also occupational and non-leisure time physical activity.

In this round of data collection also in 2003 to 2004, they had the National Cancer Institute incorporated accelerometry into the data collection. And so -- oh, I'm sorry, it should have said we had a prevalence of approximately 60 percent meeting the US physical activity guidelines of either 150 minutes per week of moderate activity or 75 minutes per week of vigorous physical activity, accumulated in 10 minute bouts. And you could see that there is a big disagreement here.

And, you know, we can quibble about this might be an unfair comparison because the accelerometer at that time

was not as sensitive as the ones that we have today, this is only a uniaxial accelerometer, so maybe if we had multiple planes it would beef up the objective. But recent studies coming out of Europe, Canada, national surveys also collecting accelerometry data have compared with self report and again they find a big disagreement between performance or objectively measured physical activity and self reported. Typically the correlations are between .2 to .3, so pretty low.

MALE VOICE: There are no error bars, maybe that's not statistically significant (laughter).

PATEL: Okay, so there are many accelerometers on the market now, they have a variety of settings and characteristics that can be examined. What is common to all of these is that there is basically a mass element that is surrounded by transducers, and once there is a force applied that is registered, the voltage displacement is registered and that generates a signal.

So what are these signals, they are referred to in the field as counts, and counts are arbitrary units that are specific to each make and model of an accelerometer. So these are proprietary units of acceleration, this is a major challenge because we can't compare across devices very easily, but they are very reliable and I'll get back

to that in a second.

So these accelerometer counts measure the frequency and intensity of acceleration in a particular plane or an axis of, you can think about a uniaxial displacing vertical, vertically. What is very nice about accelerometers is that the data are time stamped so you can look at daily variation in activity patterns. So it's a real time monitoring of behavior.

These data are accumulated over discreet time intervals that are defined by the user typically with more recent accelerometers, and these units of time are referred to as epochs. So you can collect them as frequently as every second and that is considered raw form, and so typically the newer accelerometers on the market now can collect between 10 to 100 hertz, so that's, you know, 10 to 100 signals per second of motion, or you can collect them in 15 or 30 second or 1 minute intervals. And what these accelerometers do then is that the devices then pick out the highest signal coming out of that 30 hertz sampling unit.

So there is a tradeoff there. When you shorten the epochs, you get a lot more data out of the accelerometer, but you consume more memory and you reduce battery life. But the field is rapidly evolving, the

devices are getting smaller and the capacity is improving.

There have been a number of validity studies done, and as devices are coming onto the market it seems that the correlation between the accelerometry counts are improving when measured against the oxygen consumption, VO2max studies, or indirect or direct calorimetry studies, basically correlations in the literature so far are around .5 to .9 in adults and similar in children. But like I said, there is a wide variation but that is probably due to differences in study design and the device types. I should say that these devices are very reliable, they have high inter-and-intra-model reliability.

Now I don't want to spend too much time on this, but I think it's important to kind of get and appreciate what comes out of these devices. So this is an example of three devices on a shaker. And that is going around at different speeds, different RPMs, rotations per minute. And so going from 25 all the way up to 225, what they did was this group, Chen et al at NCI, they tested these in 6 minute intervals and had 2 minute breaks in between. And as you could see, so the first accelerometer is set to collect raw data that's not processed. And so what you find is that as the rotations increase in speed you get great displacement.

Now here is an accelerometer that has filtration. So the device based filtration system is built in and as you can see already, that the filtration removes some low activity and then it also removes higher activity. So you see greater displacement in the raw data but then using the filter data we lose some information. And the reason for this is that manufacturers wanted to try to reduce error but there are physiologic movements that humans generate that we should be able to detect.

And then down here is the third accelerometer where it is also filtered but it's set to 1 minute epoch. So you're collecting all these data and averaging them over 1 minute, so that's why these are flat.

And I don't want to belabor this, but this is two accelerometers on the same person, one on the left hip and one on the right hip. This is raw data, this is filter data, and this person was asked to go through a circuit of activities and as you, just as a point of illustration, this is the vertical anterior/posterior and lateral planes, and as a person is increasing in activity going from standing, walking, running and sprinting, you see greater displacement, greater activity counts, but the filter data actually in the vertical direction is actually the signals are truncated. So this is just to illustrate that

filtration can make a difference in the types of data you get out of these devices.

Now what's typically done or what has been done is you get these data out, these are vertical axis counts across one day and again these data are time-stamped. And what people do is determine how much time is spent at different thresholds of activity. These thresholds come from treadmill tests or other objective tests where the accelerometer counts are calibrated against a particular activity. And so people will estimate how much time someone spends in moderate to vigorous physical activity relative to the overall wear time.

So it's important to note that there are commonly used cut points, but they are not necessarily consensus cut points. The NHANES data, most people use these cut points to identify sedentary lifestyle, moderate to vigorous physical activity, and then people just apply them across the board to the entire population.

But what's important to appreciate is that there is variation in the population as to how much activity someone can engage in. And we know that, for example, that VO2max or working capacity or exercise capacity decreases with advancing age. And so it's highly unlikely that you will find an 85 year old who can achieve vertical

displacement of 6,000 or more counts and so if we try to estimate how much time this older person is spending, vigorous physical activity, by applying this threshold, that might lead to an underestimation of that person's physical activity. So activity threshold counts is a huge area of investigation, and I'll come back to that in a little bit.

So some more definitions or things to consider with accelerometry is monitoring time. We can now monitor up to 30 days, but memory and wireless capacities are improving the length of time that can be sampled. Typically what a valid day of assessment is having at least 10 hours of wear time, and 60 to 70 percent of waking hours are recommended. And what is also recommended is having at least 3 or more days of activity assessment including weekdays and weekend days.

Now in terms of where you place the device, a few years ago it was widely believed that you couldn't get accurate data out of the wrist and that you can only really get good information about overall physical activity out of the hip. But actually this has changed quite rapidly.

This was a study that was done last year or published last year where they had individuals wear six accelerometers, do a circuit of different activities, and

using machine learning or pattern recognition statistics they were able to find very similar levels of accuracy in terms of classifying people on the different activity levels, although what they did report was that the hip provided statistically significant better classification than the other sites, but the difference clinically was not really that large.

I would like to point out that a lot more accelerometers are now including inclinometers into the devices. So the Actipal is one example where you can actually detect people sitting and standing, the amount of time they spent sitting or lying down versus standing or walking. And that illustrated in this printout, where yellow represents either sitting or lying down, green represents sitting up and red represents walking time.

So I just wanted to point out that as many of you probably recognize that with advancing age activity decreases. And these are recent data from the Baltimore Longitudinal Study on Aging and again, these data are time-stamped so you can really get a feel for when peak activity occurs, how long it's sustained for, so younger people peak a little bit later and they have a longer sustained activity period, whereas older participants peak slightly earlier and then they have greater fatigue and decrease in

their activity pattern.

This is a little bit of a digression but I wanted to sort of illustrate how sensitive these devices can be. Coming back to the NHANES data, these are data among older adults who do not have any self reported mobility limitations, and I wanted to show that a biological marker of hemoglobin concentration is pretty strongly associated with activity patterns. So these two bars represent individuals who are just below the WHO cut points for anemia, either half a gram to a full gram below the WHO cut point. Not severe anemia, very mild, you see that they have decreased activity patterns. People who are just below or above the threshold have intermediate activity patterns, and those who are a gram or 2 grams above the threshold have higher activity levels. So this is just to illustrate that the accelerometers are sensitive to biomarkers.

So we took advantage of these data, in the 2003-2004 data collection there was actually a supplement of questionnaire data on pain, and we were able to classify individuals into non-chronic pain, chronic regional pain, and chronic widespread pain and then examined activity patterns. And this is just an overall depiction of the results. You see that people with chronic widespread -- without chronic pain-- have higher activity patterns in

both men and women, and that those with chronic widespread pain had decreased physical activity patterns. And we looked at different activity thresholds, what we found was that there were no differences in terms of sedentary time or in lifestyle activity, the big differences were in moderate to vigorous physical activity. Typically people with chronic widespread pain had 6 to 7 minutes less fewer minutes in higher levels of activity than those without chronic pain.

The previous NHANES results are similar to a study that was published last year where fibromyalgia patients were actually diagnosed by a physician, unlike our study that relied on survey data, and again they compared patients with fibromyalgia with age matched controls. These are women and what you can see here is that these are physical activity self reported data, and these are the accelerometry level parameters that are objectively measured.

And if you look across in terms of moderate activity measured by self report, you see very low correlation that are totally nonsignificant in the fibromyalgia. In the control population you do see significant correlations between self reported moderate activity levels with accelerometry determined moderate

physical activity levels, but the correlations themselves are not very large.

So moving onto another study where they actually, or we assessed peoples' walking capacity through a 6 minute walk test, Dr. Phillips has described previously, and we had the participants, these are 319 people in a subsample of an Icelandic cohort study, and this is just to illustrate that instead of looking at distance traveled over the 6 minutes, we converted it to walking speed. And you see a very nice correlation, .8, of vertical axis counts displacement with greater walking speed. And when you look at breaking it out by vertical, anterior/posterior, mediolateral, you see strong associations as you walk in the vertical displacement, less strong association in the anterior/posterior, and very little association with mediolateral. So this is just to show that in a maximum walking test you do see strong correlations with an accelerometer detecting walking pattern.

And then these people were asked to continue wearing the device for the following week. And so when you stratify by their level of performance on the 6 minute walk test, 500 meters reflecting high functional capacity, walking capacity, and those with less than 350 showing

clinically significant decreased walking ability, you see very different activity patterns over the course of the subsequent week.

I just want to also illustrate that these accelerometers have recently been used to predict the onset of disability in basic ADLs in a cohort of older adults in Chicago. Just to illustrate they showed the hazard function for developing ADL disability over a 5 year period comparing those in the bottom tenth percentile of activity to the top tenth percentile of activity, and you can see there is a wide difference. So this provides some validity for accelerometry for predicting disability.

This was also a nice study of rheumatoid arthritis patients. I believe there were 30 of them. And what they did was they had them wear an accelerometer for 6 weeks, they had assessed disease activity through -- I'm sorry, I'm blanking on what this stands for --

MALE VOICE: Disease Activity Score.

PATEL: Thank you, the Disease Activity Score-28. This is I think the estimated sedimentation rate and they had them come back 6 months later. And so what you could see here is that there is a correlation of -.46 whereby people who actually had decreased or improvements in their disease activity had increases in their physical activity

patterns compared to those who had worsening of their disease activity pattern.

Now I just tossed this in because there was a discussion that these accelerometers are not able to detect different walking activity, standing activity, but, in fact, the field is rapidly moving and using machine learning techniques and pattern recognition. This group out of Hopkins just published a method fro identifying not only different discreet activities, but also modeling the transition between activities.

So here they had people walk for a minute or so -- no, no, for 20 seconds, and then they had them lay down, rest while lying for a few seconds, get back up and stand, and then do it again. And so what they did was they had the, this is a preliminary analysis, but what they showed was they did a, they had someone observe the participant do these activities and record their timings and then they predicted it with the accelerometry data.

Then they called them, the participants continued wearing the accelerometer at home, had them redo the protocol, and use the first visit as a training dataset and then at the second visit their algorithm for predicting standing, sitting, lying down, was very, you know, you had high classifications, this is just an N of 1. And they did

have improvements in trying to classify transitions between activities. So this is just preliminary data but there are new, there are studies, it seems like it's been coming out every week, showing that these accelerometry data can be used in a very in depth way.

So some considerations. There are a lot of pros, you get objective, continuous monitoring of free-living activity. These are very dense data. A key strength is that we could detect lighter intensity activities, and the data are collected passively. The drawbacks are the costs are more expensive than self report, but the costs are dropping down rapidly. I mean 3 to 4 years ago the prices were in the $600 range, now we're in the $200 to $300 for these research level accelerometers.

One major drawback is you lack context. You have to do some sort of training of the dataset to really be able to know what these individuals are doing. And so there's a lot of methodological work being down to try to identify patterns of activity that can be applied to participants in studies.

These accelerometers do not measure bicycling, they do not measure strength training or isometric exercise. So that's a drawback, so you underestimate some activity there. And a key problem is that there is lack of

industry standards, there's a lot of proprietary algorithms out there but there is a movement to push these companies to open up their algorithms to the research community.

And finally, these data are, you know, these are rich data, you are getting about 2 to 3 million signals, data points on each individual per day, so it's a lot of data. So it's not as simple as handling survey data.

I would also like to just mention that there's a lot of interest by consumers to start using devices to monitor themselves. The *New York Times* ran an article that was fun to read and if you go online you can actually watch videos of the reporter do activities and watch his accelerometer register those activities. And then *Forbes* ran an article about the marketplace for these devices and how it is very promising. So it's just an indication that people are accepting of these technologies.

Okay, I'll leave it there. Thank you very much.