ACTTION - IMMPACT XX - Assessment of Pain Outcomes Clinical Trials of Chronic Pelvic Pain and IBS

July 14, 2017

A Matter of Record
(301) 890-4188

ACTTION

INITIATIVE ON METHODS, MEASUREMENT, AND PAIN ASSESSMENT IN CLINICAL TRIALS (IMMPACT-XX)

Recommendations for the Assessment of Pain Outcomes in Clinical Trials of Chronic Pelvic Pain and Irritable Bowel Syndrome

Friday, July 14, 2017
8:35 a.m. to 3:20 p.m.

Westin City Center
Washington, D.C.

C O N T E N T S

P R O C E E D I N G S

(8:35 a.m.)

DR. SMITH: Good morning, everyone. We are going to get started, so if you can take your seats, that would be very helpful. Thank you.

The first speaker we are going to have this morning is Dr. Jennifer Gewandter. She is an assistant professor in the Department of Anesthesiology at the University of Rochester.

Presentation – Jennifer Gewandter

DR. GEWANDTER: Good morning, everyone. Thank you for being on time. It's very nice. This morning, I am going to be talking about a systematic review that we did looking at all of the clinical trials in the areas we have been talking about today that we found.

The objective of our systematic review was to summarize eligibility criteria and outcome measures from previous RCTs in order to inform our discussion and recommendations for future trials. When designing the coding manual for this review, we thought about a few things that we have already

covered today, that there are multiple symptoms and we have to control false positive rates. They sometimes include recurrent pain, pain affected by the other symptoms as well as activity-specific pain; many potential causes of lower abdominal pain that we have to rule out if we want to have a homogenous population, as well as this idea that we have mentioned of overlapping conditions.

For my presentation, I am just going to outline our systematic review methodology and then the characteristics of the trials that we found and then summarize the trial inclusion and exclusion criteria; the outcomes measures and endpoints, which I think Dr. Johnson did a good job of describing what the difference between those two things are; and summarize the methods that were used to adjust for multiplicity in these trials.

First, I just wanted to acknowledge that this was a group effort. A lot of you in the audience were involved in the planning as well as in the feedback stage, and this manuscript is in preparation.

Page 5

For our review, we included the conditions that we are talking about today as well as a term for chronic pelvic pain. We searched on the condition names and the word "pain." We also searched on drugs that are approved by the FDA or EMA for these conditions to see if we could find any other trials.

Inclusion criteria for the systematic review was that the trial was randomized, it was a pharmacologic treatment, it either treated one of the conditions that are listed and that we're covering today, or included patients with chronic pelvic pain and they didn't require specific etiologies. The trials had to be double-blind and have at least one pain-related outcome reported in the abstract, and this could include discomfort.

Our search resulted in 121 articles from the first search, and then two additional articles from the second search that we didn't identify in the first search.

Here is the breakdown of what we found in terms of the conditions. The majority of the

Page 6

articles reported trials for irritable bowel syndrome, and then interstitial cystitis and chronic prostatitis were the second most common. We didn't find a lot of trials that included a broad pelvic pain indication or vulvodynia.

The majority of the trials were published after 2000. And interestingly, only about 25 percent of them investigated drugs that we consider to have a putative pain mechanism, so things like opioids, anticonvulsants, antidepressants. Everything else looked at things like anticonstipation, antidiarrheal agents. A little over half were sponsored by industry.

Now I'm going to talk about the inclusion and exclusion criteria. What I'm showing here are the percentage of trials that had these inclusion criteria. The darker bar is IBS, and then the other bar is pelvic pain put together. The most common inclusion criteria was a minimum duration of pain, and then the second most common was an established diagnostic criteria. For IBS, that is the Rome criteria.

Page 7

One thing that I noted that was interesting was that fewer than half the trials in both the IBS and the pelvic pain group included a minimum severity of pain, and that is something that in the pain conditions that I think about on a daily basis, we would always include in a trial. I'm blanking on what they are because of the dot, dot, dot. It's a minimum score on a composite measure, so like the prostatitis composite score was inclusion criteria for the pelvic pain trials.

The second-to-last one is diagnosis by a clinician without any kind of definition really, and then the third one was some kind of imaging. So this was common in things like interstitial cystitis that Dr. Lai talked about yesterday.

Is there any way I can stop for a second? Sorry.

(Pause.)

DR. GEWANDTER: Thank you for your patience.

For exclusion criteria, the most common exclusion criteria was a comorbid condition that could be associated with abdominal pain. This

Page 8

ranged from things like a UTI to a kidney stone to cancer. The next most common was an imaging or exam or lab testing to try to identify these things, so a test for infection.

A lot of trials prohibited use of certain drugs, which I will talk about the different types in a minute; abdominal surgery, alcohol or drug use, as well as psychiatric conditions.

The drugs that were prohibited, opioids were commonly prohibited. Then often, there was a phrase that said "treatments for the condition," but it wasn't specified what that meant; antidepressants, anti-inflammatories. Some trials just stated all analgesics were excluded. Hormones were excluded sometimes, and anticonvulsants.

Now I am going to change to primary outcome measures and endpoints. Just so we are on the same page about the denominator in my summaries, 86 trials or 69 percent identified one or multiple primary outcome measures. The others just didn't identify one. An example of an outcome measure is the 0 to 10 pain numerical rating scale. That is

Page 9

what I am talking about when I say a measure.
Then 67 or 54 percent of the trials identified a single primary endpoint. What I mean by that is something like response defined as 30 percent improvement in pain intensity at trial endpoint. These numbers are what's used for the denominator and the percentages in my next set of slides.
The most common primary outcome measure -- there really weren't that many commonalities, I think, as one takeaway from these slides -- was a composite pain and non-pain outcome measure, which was used a lot in the IC and prostatitis studies; this idea of an overall symptom relief that was specific to the disease, so please rate your IBS symptom relief.
IBS or abdominal pain and discomfort relief was common, and then less common was just a measure of pain intensity. Sometimes people identified multiple primary outcome measures, one of which was pain intensity. Then the next most common was a symptom relief question that was not specific to

Page 10

disease, so just please rate your symptom relief in general.
I also summarized the non-primary outcome measures. Interestingly, pain intensity was very common for a non-primary outcome measure and also diaries of either non-pain symptoms and signs, so like number of urinations, number of defecations.
Quality-of-life measures that are specific to the disease were included frequently, as well as measures of depression, anxiety, and quality of life that was not specific to the disease, so something like the SF-36.
Now I want to cover how the articles turned those measures into endpoints. I presented this a little bit differently, mainly because if you just take a quick look at the numbers, there is not a lot of commonality between the trials. So what I've done here instead is group them into families.
The first family are responder endpoints that are based on over a certain percentage of time, so how the patient is responding in, let's say, 6 out of 12 weeks, for example. The things

Page 11

that people use for defining response was adequate pain relief for a certain percentage of time, adequate IBS symptom relief for a certain percentage of time, adequate pain relief and improved bowel movements for a certain percentage of time, and that would include the definition by the FDA, like the FDA guidance would be included in there, and adequate improvement in stool consistency over a certain percentage of time.
Then there were also response endpoints that were based on a single time point, so for example, just the endpoint week, so adequate symptom relief at endpoint, adequate improvement in pain and non-pain composite outcome measure at endpoint, and adequate improvement in stool consistency at endpoint.
Then there were the severity endpoints, so just comparing. This would be like a T-test, continuous outcome measures, comparing the severity or change from baseline in pain at endpoint; the severity or change from baseline in pain and non-pain at composite at endpoint; and again, stool

Page 12

consistency or constipation at endpoint.
Also, a couple of the trials -- there is a typo there; sorry, that should be a zero for the pelvic pain -- included a biomarker endpoint, so that was interesting. Again, these are primary endpoints.
Then there were a couple miscellaneous that were like a model that incorporates a bunch of different times over the study like repeated measures ANOVA. Then there were a couple single-dose studies, so summary of change in pain intensity at specific time after receiving a dose of the treatment. And I just put the other in for completion, but if they only occurred in one trial, I did not summarize them here.
Methods to adjust for multiplicity, the endpoints that I just talked about, some of them can combine two symptoms into one outcome in order to incorporate two symptoms without inflating your false positive rate or type 1 error rate, but you can also do this statistically by adjusting for multiplicity.

Page 13

So 71 or 57 percent of the articles did not identify a primary analysis. This would be not only identifying the primary endpoint, but then also describing how you were going to do the statistical analysis in sufficient detail. Thirty-five percent of the articles identified one primary analysis, and 7 percent identified multiple primary analyses, and of those 9, 7 adjusted for multiplicity.

The methods that were used were primarily a gatekeeping strategy. Forgive me if I'm boring you and you already know this, but gatekeeping is when you have two primary outcome measures but you give then an order. Let's say the first one would be pain, and you do an analysis on the pain outcome. If it's positive, then you can do an analysis on the constipation outcome, and your significance level could be 0.05 for both of those analyses. But you wouldn't move forward if the pain outcome was not positive.

Then one trial used Bonferroni correction, which is when you split the alpha between two

Page 14

analyses, so you have to get a 0.025 for both analyses for the trials to be considered positive -- either one can -- sorry -- hit 0.025.

Just in conclusion, our review identified high variability in entry criterion outcome measures even within these end-organ conditions. There were deficiencies in identifying single primary analyses or adjusting for multiplicity in the articles. But they did give us multiple examples of methods to combine symptoms into single endpoints or adjust for multiplicity; again, these responder definitions using different baseline, different time frames within the trial like over the whole trial or at the endpoint, composite outcome measures as well as gatekeeping and Bonferroni approaches.

For the purposes of our discussion later this afternoon, I just wanted to quickly mention some of the methods that you can use to adjust for multiplicity, or combine outcome measures, or combine endpoints that weren't covered in the systematic review.

Page 15

One of them was mentioned yesterday, this idea of co-primary analyses. You do an analysis for pain and an analysis for constipation. They both have to hit at 0.05 for your trial to be considered a positive trial or note that the treatment was effective.

There are stepwise procedures that are like a Bonferroni correction, but they are a little bit less strict. For example, Holm where let's say you have two outcomes, you do an analysis on both those outcomes. As long as one of them hits a p-value of 0.025, the next one can hit a p-value of 0.05 and you can still consider the trial positive.

Then finally, there is this relatively new methods that rank participants based on their combined treatment response on multiple outcomes. An example is DOOR, which we distributed an article on this by Scott Evans. I am just going to try to explain it. I am not a statistician, but I just want to give you the 30,000-foot view of this.

An example of this would be if you want to incorporate in your endpoint how patients respond

Page 16

to pain but also whether they take rescue medication or not. One of the main advantages of this type of analysis is you can incorporate competing interests. If you take rescue medication, your pain might look better, but that might not necessarily mean you're a better responder because you've taken rescue medication.

In order to do this analysis, you rank participants. This might be an example ranking scheme. Patients who improve by greater than 50 percent on their pain and they take no rescue medication, that's the best outcome. Then the next outcome would be that they improve by 50 percent, but they took rescue medication greater than 20 percent of the days.

Then the next would be they have less than a 50 percent improvement in pain, but they don't take any rescue medication. Then finally, they have less than a 50 percent improvement in pain, and they also were taking a bunch of rescue medication.

You can obviously use finer gradation for this, and you might not necessarily agree with the

Page 17

order that I put here, which is one of the challenges of the method. So you rank patients based on these criteria. And what the DOOR probability tells you, when you do the analysis, is the probability that a randomly selected patient in arm A has a more desirable outcome than a patient in the control arm.

The advantages of this method are that it uses outcomes to analyze the overall patient experience rather than patients to analyze each individual endpoint. When you do a co-primary analysis, you might show that, overall, people have improved pain and, overall, they have improved constipation. But the patients who improved in pain could be completely different than the patients who improved in constipation, so you don't really know what their overall experience is.

It has this appealing probability interpretation that we usually can't do with frequent [ph] statistics that people like. And again, it deals with this competing outcomes issue of if I take more rescue, my pain will be lower,

Page 18

but that doesn't necessarily mean the drug was better.

It may have more power than a dichotomous composite responder analysis. The responder analysis that the IBS guidance gives us where you have to be a responder, you have to improve 50 percent on pain and somewhat on the constipation scale, or the stool consistency scale, because that's just a dichotomous analysis, this might have more power than that.

The limitations are developing that ranking scheme -- you know I made up -- so how much input do you need from patients, how do you decide what those ranks are. Also, just like any composite measure, the differences could be driven by a single measure. It could be all driven by if you did this for pain and constipation. The change in the probability could be all driven by pain, but constipation could have not changed. But that is true for any composite.

With that, I will thank everyone again who was involved in the systematic review, as well as

Page 19

Dr. Evans for reviewing my slides and hopefully preventing me from embarrassing myself.

(Applause.)

DR. SMITH: All right. So next, we have Dr. Quentin Clemens, who is a professor of urology at the University of Michigan Medical Center.

Presentation – Quentin Clemens

DR. CLEMENS: Thank you, and I have really enjoyed the discussion and meeting so far. There was a lot of talk yesterday about the MAPP, so I just want to bring everyone up to speed about the organization and some of the main findings.

A couple points about the title slide, we have two Ps, so we are part of the club.

(Laughter.)

DR. CLEMENS: The other is that as someone who is involved every day with the data analyses, I always feel tempted to put the title being questions unearthed by the MAPP rather than lessons learned because it is very tempting as you are always thinking of the weaknesses of what you found or what is the next analyses. But I will do my

Page 20

best to summarize what we have found to be the main findings.

MAPP stands for Multidisciplinary Approach to the Study of Chronic Pelvic Pain. Someone asked me Friday evening what does MAPP stand for, and even I got it wrong because I put a urologic in there. This is the official title, but we just refer to it as MAPP.

It is funded by the NIDDK, and we are dedicated to studying IC and chronic prostatitis, and we have coined a term, "urologic chronic pelvic pain syndrome," to encapsulate both of those conditions.

A little editorializing here, when a man has pain essentially from the nipples to the knees and it is not associated with bowel movements, they tend to come to see urology, and often they get diagnosed with chronic prostatitis, chronic pelvic pain syndrome. And that is not inappropriate necessarily. We're used to it.

Women are different. When they have pelvic pain, they see a gynecologist. And if the pain

Page 21

tends to be focused on the bladder, then they ultimately sometimes find their way to urologists and get diagnosed with IC.

So we are combining these two conditions, but they are actually not the same. There is a whole population of women out there who have pelvic pain. It's not endometriosis necessarily, and it is not associated with the bladder. We are not studying them in this.

For instance, you would expect that if we compare men and women using these criteria, that the women will have more bladder symptoms because that is the definition of IC, and what we are seeing, that's what we found. As we noticed yesterday, we were surprised a little by how many bladder symptoms men have, but I think as we think about what we are studying here, I think keeping in mind that these are a little bit apples and oranges is useful.

Why do we need MAPP? Well, we haven't made much progress in helping these patients. We in urology and urogynecology had not, before the MAPP,

Page 22

really worked closely with smart people like yourselves. There is more and more of a feeling that these patients represent probably a multiplicity of different etiologies; in other words, there is a need for phenotyping. And hopefully if we get a better understanding of subgroups, that will lead to more targeted therapies and better outcomes.

It is organized. There are six main what we call discovery sites listed here. One of the smartest things the NIDDK did is they required that each site have a non-urology investigator as a co-PI. I work with Don Clauw. UCLA has Emeran Mayer, et cetera, et cetera; Dedra Buchwald from University of Washington. That has been for me the best experiences about this, and it brings more energy and more insight, so that's been really good.

Then we have some specialized discovery sites that don't recruit patients but conduct some other ancillary studies; the data coordinating center, which Dick Landis here runs, and a tissue

Page 23

and technology core at University of Colorado; then of course the NIDDK. And there is an oversight executive committee. Mike Pontari is a member of that.

Here is a nice map that shows we have a fairly decent geographical representation, including Canada.

The goals are to better understand the treated natural history of UCPPS; identify clinical and research factors that hopefully will define relevant subgroups, which can inform future clinical trials and address underlying disease pathophysiology.

Our inclusion criteria were really quite broad. They had to have a clinical diagnosis of IC or CP. I think that is important. There were some patients that found their way to us that maybe saw an ad, and so we made allowances to say, well, they tell us they were diagnosed with IC.

So what we made sure is that there was some clinical evaluation done when they came to the initial appointment by a clinician, just talked to

Page 24

them a bit about their symptoms and make sure it wasn't obvious they had endometriosis or something along those lines, as opposed to some idea where we would advertise widely and anyone with the right types of symptoms could get in. We wanted to be sure as best we could that these really were IC and CP patients.

We talked about exclusions a bit yesterday. These are pretty standard across trials and studies.

We did want to examine whether those with a short duration of symptoms were different, those with longer, so we oversampled for subjects with less than 2 years of symptoms. That is what we defined as early.

Now, to cut to the chase, it didn't really matter very much in the analyses we've done, so we haven't followed up much with that in MAPP II, but it turns out that the patients with short durations of symptoms really tended to not look much different at all than those with longer duration.

Then we had two control groups. One were

Page 25

those with no urologic symptoms at all, and the other was patients diagnosed with fibromyalgia, IBS, or chronic fatigue. The RFA specifically focused on those three conditions, so we focused on those three conditions.

We did not recruit intentionally people with migraine or other conditions that are typically part of that chronic overlapping pain condition group. This was our positive control group.

You-all, I'm sure, are very interested. I didn't list all the questionnaires because most of the audiences don't care too much about them, but all kinds of different, psychosocial symptoms, catastrophizing, IPIP questionnaire, et cetera, et cetera. This was about a 2 to 3-hour battery of questionnaires that were administered, a lot of details about their urologic symptoms, of course, psychosocial symptoms, pain symptoms in general, the body map.

The physical exam was fairly minimal. In MAPP I, we asked do they have pelvic tenderness, pelvic muscle tenderness, yes or no. In MAPP II,

Page 26

we are doing more of a detailed, which pelvic muscles are tender, and also importantly, does the exam reproduce at least some of your symptoms to try to get a little more detail about that.

We obtain bio specimens. We did neuroimaging and QST, and I will talk about those a bit at the end. It took some time to get those all up and running for a variety of factors. I don't think there had been a multi-institutional group like this who had ever done neuroimaging before. So it's always been one-off. One site does something; another site does something.

So we get everyone together, agree on a protocol, make sure all the scanners were equilibrated equally, et cetera.

As a result, the number of subjects who have the questionnaire data, the QST, because that took some time, and the neuroimaging, when you do that Venn diagram, it's actually pretty small for MAPP I.

In MAPP II, now we're halfway done with recruitment. We already have way more patients who

Page 27

have had all three together than we do in MAPP I. So there is an example of some of the advances you can make with continuing things.

This is the flow. The subjects were recruited, as I mentioned. All of them, including the controls, of course, did the baseline phenotyping that I just described. Then for the controls, they were done. Then the UCPPS patients were then followed for a year. They came back at 6 months and 12 months and had pretty much the same assessment except no QST or neuroimaging. That was just done at baseline.

Also, throughout the year, they underwent biweekly internet assessments. So they were paid about $5 to do that. And I'll show you, but they really were very compliant with that. So we have a huge amount of data, a lot of repeat measures, et cetera.

Then people who were in the study then were eligible to have site-specific studies done, kind of as add-ons, based on the interest of the various sites. Importantly, there on the right, the

Page 28

regular treatments were allowed. We tracked that to some degree. Every 2 months, we assessed what treatments they were currently undergoing, so we had some idea of the treatments, but not super closely following that.

This is a treated natural history study. This is sites that we think know what we are doing in terms of treating this, and so there are a variety of different treatments that were prescribed. So when we talk about someone who got better, they got better based on probably the treatments they received.

We recruited overall 424 UCPPS patients, 415 healthy controls, and 200 of the positive controls. As you can see, the positive controls were mostly women just because they tend to have those conditions more commonly.

The first point is that we found using baseline data that our MAPP subjects look similar to those that were previously reported in the literature. Mean symptom duration, 8 to 9 years. We used some of the symptom scores that could be

Page 29

compared with older studies as well and looked very similar.

The other here is 83 percent missed no more than 3 of the 26 planned contacts throughout the year. So it's really a tribute to the patients and also speaks to how desperate they are to find better treatments. They are very willing to bend over backwards for us to help.

A couple of the themes that have emerged as being as important: The one is the degree of widespreadness of pain is important. Here we're finding body maps to be increasingly useful, so if we define pelvic pain only as those three regions there, we understand that the IBS people are going to say, wait a minute, that's our area, too.

We are looking into that more, and in MAPP II, we have actually divided the abdomen into some different quadrants to try to help. We also have the CMSI instrument, which has a module for IBS. So if bowel symptoms are reported, then there is a separate model triggered to really go through diagnostic criteria. So we have those data that we

Page 30

can look at.

This was essentially a baseline tool. We have found that those who have more widespread symptoms have worse urologic pelvic pain. In MAPP II, we have repeated measures for this, including during a run-in period to look at the stability and help to define the phenotype maybe better at baseline. Also, we have severity as a measure. This doesn't.

So ultimately, maybe if they have trivial head pain, we might exclude them as having pelvic pain and beyond, for instance. So we're trying to look into this in more depth, and we are looking at it in more depth.

In terms of the psychosocial symptoms, our urologic patients are every bit as affected in this regard as fibromyalgia, IBS, chronic fatigue patients. As you might expect, if you have these chronic fatigue syndrome, et cetera, symptoms, you are more worse off. You are worse off, worse quality of life, worse psychosocial symptoms.

We found about 40 percent of the females and

Page 31

30 percent of the male subjects had one of these diagnoses. And then of course it goes up if you add migraine and other types of overlapping pain conditions. This has been reported previously for women, not so much for men, so it's somewhat novel data, examining this as closely as we did for the men.

We discovered what we call bladder sensitivity phenotype, and this was briefly mentioned yesterday. The first point is that men had more bladder symptoms than we thought. This doesn't mean that all the men have IC, but what it does mean is we should pay attention to that when seeing the patients because it does seem to correlate with a worse quality of life and would suggest if we helped to address those bladder symptoms -- or I guess the other way to say it, if we ignore the bladder symptoms, which I think is perhaps what is done not too rarely, we won't be able to help them as much. And this bladder hypersensitivity seemed to be associated at baseline with worse quality of life and more severe

Page 32

symptoms overall.

Jamie Griffith, who's a psychometrician at Northwestern, led this where we basically looked at unstructured factor analysis at baseline of the symptoms and found that two factors emerged: pain and urinary symptoms. This was similar in men and woman. Then we also looked longitudinally and found that not only did they look different at baseline but they tracked differently. So this was the subject of a good bit of discussion yesterday.

To date, a lot of the outcomes for these trials have been composite scores for urinary and pain, and so what this leads to is a conclusion that we probably should have pain outcomes and urinary outcomes separately.

In fact, John Farrar has led an activity, grant that has been written up to try to retrospectively look back at the existing clinical trials, try to separate, the best we can, men and women into pain or urinary phenotype, and look at the types of treatments they get, and see if by doing that -- and in having pain and urinary

Page 33

outcomes, which we can derive from the trial data, see if we can examine this concept using existing data to see if it pans out that perhaps some of these negative studies might look positive if we subcategorize them like we are proposing in MAPP.

When we looked at longitudinal data for whether patients get better or worse, the first concept was to just look at the slope of the symptoms. This is a pretty stable of group of patients that don't tend to change their symptoms a heck of a lot over time. So using a slope, most everyone just ended up looking stable.

We looked more closely into the data and came up with this functional clustering algorithm. And these next two slides are the ones where I am happiest that Dick is here so that if there are questions later, he can go over exactly how this was done.

We still, as you can see, had 60 percent who were in that stable group, but we had 20 percent who were improving and about 20 percent who were worsening over time. We did this for both the pain

Page 34

and the urinary symptoms. We have composite scores using the GUPI questionnaire and the IC symptom index to define the pain symptoms and the urinary symptoms. Then once we had defined these variables, then we could examine predictors of who gets better, who gets worse.

Another editorial comment. This, Dick, was about six months of work, right? It took a while. When we do clinical trials, we prespecify an outcome, and then we get to that point, and then we look at it, and then we are done.

These types of cohort studies are much different, and in my opinion much more difficult to run because you are constantly reassessing as you go. And it seems pretty simple that you ahead of time say, well, these are some variables we hypothesize will correlate with improvement.

You can even say we are going to measure improvement one way or the other, but then as you get into it and have all the data, you say, well, there's probably, with all this data, a better way to measure improvement or worsening. You don't

Page 35

necessarily want to publish a very simplistic paper that concludes one thing, and then actually have a more detailed analysis which you think is better and concludes something else.

So then you kind of put the brakes on things, spend six months or so to define what it means to improve or get worse, which is kind of a fundamental component, and then you can publish your paper, and then move on using that variable longitudinal data for other analyses.

This paper, as you can see, 2017, it just came out about a month ago. We are seven or eight years in. I don't think we have been resting our feet the whole time. These things take a lot of time.

The predictors of better outcomes included -- and the most important one is the higher baseline symptom severity. Other predictors were less widespread; pain, less; non-urologic symptoms based on the CMSI and body map; better overall physical health and mental health; with the measures you can see here, sleep and fatigue.

Page 36

The mental health particularly -- and this has been shown before -- catastrophizing was important and also perceived stress. Some of the factors that were not important were age, sex, symptom duration, and perhaps somewhat surprisingly, anxiety and depression.

As we do more and more of these analyses, we find that sex typically washes out. So we certainly acknowledge that there are differences between the sexes in the types of symptoms that they often present with, but if you actually have the same symptoms in a man and woman, the sex doesn't matter, and that's what we've found repeatedly.

That's one of the reasons for this rationale or this UCPPS nomenclature because sex doesn't seem to matter as much as perhaps was thought. And I've already mentioned symptom duration has not really panned out as being very important, at least as we defined at as two years.

We talked about flares yesterday. This every 2-week assessment included a question, have

Page 37

you had a flare in the last 2 weeks? Before we did that, we did focus groups that showed that when we asked about flares, the patients understand what we're talking about, so that was reassuring.

There's one paper that's been published for women, another that's in the works for men, with the results of the focus group analyses. That's where we learned that some patients have flares that are minutes in length, et cetera.

Women have more flares than men. Ninety-five percent of the cohort report at least one flare, and you can see the distribution here with 40 percent reporting 10 or more flares. This was more common with individuals who had widespread pain and those who had more severe bladder symptoms.

The other interesting thing we did was when they had a flare twice, it triggered a flare supplemental questionnaire: In the last two weeks, what foods have you eaten, what sexual activities, what exercises, et cetera? Then there was also two times when they said, no, I didn't have a flare,

Page 38

that same supplemental questionnaire was triggered. So we had an internal control, if you will, within the patients.

We didn't really identify dietary factors or much in the way of activities that seemed to trigger, using those methods. There were some question of maybe having a preceding UTI. And one of the things that's led us to do in MAPP II is to look more closely using mobile apps at some of these flares that may be more short term and see if there's something we can learn from that since we didn't identify clearly any risk factors across the group for flares in MAPP I.

I mentioned that we had many, many observations here, so one thing we looked at and demonstrated, not surprisingly, is that there's a significant regression to the mean effect.

In MAPP II, we're incorporating a four-week run-in. And this doesn't really apply too much for clinical trials, I suppose, but certainly for cohort studies, having a run-in period and then setting the baseline after 4 weeks, that's

Page 39

something we've incorporated in all of our data analyses. You can see that if you don't account for the run-in period, the number of people assigned to different categories of improved, worse, and change, it changes to some degree.

We also looked at variability. We have every 2 weeks, and you can look at how their slope is or how they do over time. You can also look at the volatility of their symptoms, and we can assign a high, low, or medium variability group.

We are looking at, for instance, healthcare seeking. I don't have a slide on that, but every 2 weeks, we ask them did you go to the ER, did you go to see your doctor for your symptoms. We can look to quantify the degree of healthcare seeking and correlate that with various things, including symptom variability.

We concluded that the phenotyping should focus on pain localization, pain outside of the pelvis, the presence of chronic overlapping pain conditions, and bladder hypersensitivity. We should not use a total symptom score. We should

Page 40

have pain and urinary separate.

Very briefly, we talked a little bit about pain testing. It is nothing like this. It uses a device like this where there is pressure put on the thumb bed, and then the subject basically says now it's starting to hurt. And now we know this is about as much as I can tolerate, and you can generate curves and compare them across different groups.

It has been demonstrated -- this is a measure of global hypersensitivity. Our urologic patients are just as sensitive as fibromyalgia patients, et cetera, on the global level. It's interesting, when you measure that as baseline, that does seem to associate with some longitudinal outcomes like number of flares and likelihood of improvement.

Then the neural imaging, again, not necessarily relevant for clinical trials, but very briefly, we can see at least at 3 months, there are certain resting state neuroimaging findings that seem to correlate with 3-month outcomes, so that's

Page 41

interesting.
Other methods have shown that our patients seem to have an increased signal in the area of the pelvic floor, which is really cool because that correlates with what we see clinically. What we're wanting to do is in MAPP II, as I mentioned, we're being more detailed about the pelvic-floor exam and seeing if there's some correlation with those who have pelvic tenderness that reproduces their symptoms, do you get maybe even a better signal?
This just demonstrates that there is similarities between our patients who have widespread pain and fibromyalgia patients, who by definition have widespread pain. So we're seeing the same types of signals using these neuroimaging techniques.
In the second phase, now a couple things we're doing, we're following the patients for 3 years instead of 1 year. We're following them a little less frequently. We're getting longitudinal neuroimaging and sensory testing. In MAPP I, as I mentioned, we only did it once. We're following up

Page 42

on some certain biomarkers. I didn't talk about biomarkers here because that's not really necessarily relevant to a clinical trial.
Very importantly, we're really focusing on treatments. We're tracking their treatments monthly, and we're really wanting to correlate our phenotyping with treatment response, not by assigning a treatment but by following them closely, having them contact us when there's a treatment change, again, prospectively following them monthly.
Ultimately, the question here is can we identify a signal that may be widespread pain patients seem to do a little better with tricyclics or something. That may help us then organize and set up clinical trials on a small scale for the next phase.
Here's the website. Thanks for your attention.
(Applause.)
DR. SMITH: We're ahead of schedule actually, so if there are any very specific

Page 43

questions that you'd like to ask Dr. Clemens, we can do that now.
DR. KATZ: Hello.
DR. CLEMENS: Hi.
DR. KATZ: How did you define the centralized phenotype exactly?
DR. CLEMENS: Well, the main way has been with the body map using the number of sites that -- and we're, again, continuing to improve that definition, but that's the way though. Pain in the pelvis only versus pelvic pain and beyond.
DR. KATZ: Was there a specific criteria? How many sites, how many body sites did the patients have to endorse before they were considered centralized?
DR. CLEMENS: We've evaluated that in different ways. I think Dick -- 3 or 4 sites total, so we had to look at a gradient, but I think it was not just one single site. I think it was three total outside the pelvis.
DR. KATZ: What do you think is the best way of determining --

Page 44

DR. LANDIS: I don't know if Henry is still here or not, but the paper that just came out, we had an intermediate group where they have one to 2 sites beyond the pelvis, and then 3 or more was the basic gradient from none to intermediate to widespread. That gradient tracked quite strikingly with many different symptoms.
DR. CLEMENS: We are, in MAPP II, as I had mentioned, looking at severity. I know John Farrar has been involved with this quite a bit. Even things like, well, if they have 2 sites outside, but one of them is upper thigh and one is lower thigh or something, then probably saying, well, that probably is the same thing, so even looking specifically to really try to incorporate not just severity but also is it really one area or not.
The ultimate question is what is the least detailed body map that we can get with, research and clinical use. So what you do is you look at all the data and be super complicated, and then do a pretty simple one and see how much more you get for the complicated one.

Page 45

So MAPP I was a fairly straightforward one, and those are the data that I'm presenting. In a year or so, we may have a little different recommendation, but my goal ultimately is to have a fairly straightforward body map for clinician use that can maybe be part of the minimal dataset that we propose for clinicians.

DR. KATZ: It seems to me that another question is what is the validity of any cutoff that you would choose. In other words, it could be arbitrary; well, do 3 sections or 5 sections or whatever, or you can say, well, what is the definition that means something in terms of maybe whatever.

DR. CLEMENS: Yes. So we've shown that when we talk about validity, usually that means how does it correlate with various clinical parameters. So we've shown that it seems to be important for longitudinal whether a patient gets better or worse, if it's predictive of that.

MAPP II will have 3-year data and may be able to similarly conclude that's what we'd like to

Page 46

be able to say is, yes, it matters. It will correlate with how well in general patients do with treatment. So I think that's where -- if it ends up not correlating at all, then we probably wouldn't propose it because, to your point, other than for research purposes, it really has no utility, perhaps.

John, anything?

DR. FARRAR: Nat, your question is very reasonable. In MAPP I, there is very little QST data, but in MAPP II, as Quentin was saying, there's going to be a much higher correlation with QST and these other symptoms. We'll be able to answer that question more specifically, but we don't have that data currently.

The definition of centralization, if you like, is simply a clinical definition based on the widespreadness of pain, and we understand that that's not an appropriate definition of wind-up, and centralization, and all of the things that we would normally think about in the experimental paradigm.

Page 47

What we hope is then -- well, what we will be able to look at in MAPP II once we have all the data is to look at the correlation between the widespreadness and QST and neuroimaging studies in a much more concise way to try and get at some of those issues.

DR. SMITH: I think Ralf had one question.

DR. BARON: This was exactly my question, of the correlation of QST and the body maps. But the only QST measure you did was pressure pain tolerance at the thumb; is that correct? Are there any other QST measures planned in the next MAPP II or something?

DR. CLEMENS: Yes. So we are doing auditory and visual sensitivity in MAPP II at multiple time points as well, but MAPP I was just the thumb pressure.

The other point from validity I'd make is that Bruce Naliboff did look at the correlation between the body map findings and the CMSI, and they correlated very highly when the CMSI was in the last year.

Page 48

Also the CMSI asked in your lifetime, have you had these. That didn't correlate at all. But I think that was also reassuring that, A, you can use either. They seem to be measuring the same thing, and we're focusing on the body map because it seems to be a more useful perhaps clinical tool, quicker, too, for the patient to complete.

DR. SMITH: Chris and then we'll cut the current questions unless anyone has something very specific.

MS. VEASLEY: Chris Veasley. So we've grappled with this idea of data analysis understanding that cross-sectional studies are not a great way to look at it, and then we obviously need prospective.

The problem is, is that a person can be categorized in year 1 as just having IC and just having pain in the pelvis, but in year 3 or 4, they could transition into another group.

So I guess my question is around data analysis, is there a plan to go back and look at those two time points to when a person may have

Page 49

transitioned from just IC to multiple conditions, or changing their allocation in terms of what group they fit into in the second data analysis?

DR. CLEMENS: Yes. So we are using the run-in period to establish short-term stability in working to get rid of perhaps background noise and better identify the phenotype. Then for sure, in MAPP II, they do the map at multiple time points throughout the three years, so we can look at that. And for those who have been in MAPP the whole -- so not everyone has, but certainly, we can look all the way back there and see.

There were some talk yesterday about this progression. To date, this idea that pain for IC starts in the pelvis and moves elsewhere hasn't really panned out. People still talk about it. I know Dan Clauw has this theory that it's really one disease. In some people, it starts in the head and moves to the pelvis, and others.

At least from the analyses we've done, that seems to be somewhat true where it's semi-random that the head is a little bit early; fibromyalgia,

Page 50

a little bit later. So I think it will be interesting to examine that in more detail throughout at least the three years.

DR. SMITH: Was your question very specifically for Dr. Clemens about MAPP?

DR. WESSELMANN: Yes. My question was actually in the same direction as what Chris asked. For instance, Jack Warren has published a series of papers on this topic specifically related to IC in a prospective study where the symptoms often start quite early on or can be triggered by surgical interventions in the pelvic area and then move on to widespread pain.

DR. SMITH: Next, I'd like to welcome Dr. Stephen Coons. He's the executive director of the Patient-Reported Outcome Consortium at the Clinical Path Institute.

Presentation – Stephen Coons

DR. COONS: Good morning, and thank you for inviting me. I appreciate the planning committee extending an invitation. I'm honored to be a part of this IMMPACT XX.

Page 51

I am going to talk about this journey that we've been on, lessons learned along the path to qualification of an IBS outcome measure. My footnote is that we haven't reached the destination yet. We do not have a qualified measure or measures in this case for IBS.

I'm going to talk about qualitative research that we have done in terms of concept elicitation and cognitive interviews with our draft measures. We have ongoing at the present time a quantitative pilot study in 315 patients, and I don't have data. I should have had data by now.

One of the problems that happens sometimes is things don't go as planned, as you can imagine, and we are deploying this instrument on an electronic data capture device, essentially a handheld device.

We should have had all of our data collected by now, but some of the data collection was over the period of time in which we changed to daylight savings time. And it ended up that the devices weren't programmed properly to take into

Page 52

consideration the fact that there was one less hour in the day. So some of the instruments that we were implementing to look at construct validity didn't get administered. So I only, as I say, have qualitative data to talk to you about today.

First, I want to talk about the context in which we're doing this work, and that's the Critical Path Institute. C-PATH was established in 2005 by the University of Arizona and FDA's Center for Drug Evaluation and Research, and it's a public-private partnership. It's an independent nonprofit organization, and part of our funding does still come from FDA. Most of my salary comes from this grant, so I'm very appreciative of this.

C-PATH is dedicated to implementing FDA's Critical Path Initiative by providing a neutral, precompetitive venue for collaboration aimed at accelerating development of safe and effective medical products.

Then within C-PATH, the Patient Reported Outcome Consortium was established. And we have right now about 14 different consortia, and the PRO

Page 53

Consortium is one of them, formed in late 2008 by C-PATH in cooperation with, again, FDA's Center for Drug Evaluation and Research and the pharmaceutical industry.

Our membership is pharmaceutical firms. We have 26 members, and then we have other participants, representatives of FDA, NIH, and at times, EMA. Then we have other clinical consultants, patients, academic researchers, and CROs that partner with us in the development of PRO measures and other clinical outcome assessment tools. This is a list of our current 26 members.

The PRO Consortium mission is to establish and maintain a collaborative framework with appropriate stakeholders for the qualification -- and I'm going to talk further about qualification -- of patient-reported outcome instruments and other clinical outcome assessment tools that will be publicly available. That's part of the process or part of the outcome of this is that these instruments will be publicly available for use in clinical trials where clinical outcome

Page 54

assessment-based endpoints are used to support product labeling claims.

Our goals within the PRO Consortium are to enable precompetitive collaboration that includes FDA input along the way and expertise; develop and obtain FDA qualification for PRO measures and other COA tools; avoid development of multiple endpoint measures for the same purpose.

That really is a major goal, and it's certainly not -- we haven't achieved that in all circumstances because a lot of individual companies are still developing their own measures, but to some extent, we have been able to avoid it within the context of the working groups that I'll mention just briefly.

Show the cost of developing new endpoint measures. For those of you that have ever developed a PRO measure or other clinical outcome assessment tools, it can be very expensive, a million to $2 million to develop an instrument. So we're able to share the costs across the sponsoring firms.

Page 55

Then a major goal is to facilitate FDA's review of medical products by standardizing COA-based endpoint measures that will be, as I said, publicly available. And we hope there will be uptake within the industry to use those in their trials.

Dr. Kovacs mentioned this briefly yesterday. This is the DDT, drug development tool guidance. This is talking about the qualification process for drug development tools, COA tools, clinical outcome assessment tools being one of those. The intent of that is to expedite development of publicly available drug development tools that can be widely used in drug development.

The definition of qualification is that qualification is based on an FDA review of evidence that supports the conclusion that within the specified or stated context of use, the drug development tool can be relied upon to have a specific interpretation or application in drug development and regulatory review.

Our working groups, there are 10 of them,

Page 56

and you can see the irritable bowel syndrome working group is one of them. We have an annual membership fee, and then pharmaceutical firms can opt into working groups. Indeed, then that subset of the pharmaceutical firm members then sponsor the activities that go on in those working groups. And you can see that we have from 2 to 10 firms sponsoring each of our 10 working groups.

The goal of the working groups is to produce and/or compile the necessary evidence to enable new or existing COAs to be qualified by the FDA. We don't only want to develop new measures. We would love to leverage measures that are out there and either adapt them, modify them, or use them and see what evidence is available for them, and ultimately develop a qualification package that we can submit to the FDA. But most of the instruments that we're working on now were developed de novo within the context of our working groups.

Then again, Dr. Kovacs mentioned this yesterday, in terms of the different types of clinical outcome assessment tools, and our working

Page 57

groups are working in all of these except right now, clinician-reported outcome measures. We're not moving forward right now with any ClinRO measures for qualification.

The IBS working group was established in March of 2009, so we've been working on this for quite a while; three pharmaceutical industry sponsors, Allergan, Ironwood, and Takeda.

RTI Health Solutions was selected as the working group's contract research partner, and the specific goal was to develop and obtain FDA qualification of three patient-reported outcome measures of the signs and symptoms of IBS-C, IBS-D, and IBS-M for use in assessing primary endpoints in clinical trials to establish treatment benefit.

Much of what I'm going to talk about today is discussed in this article that appeared relatively recently in Value in Health, development of the diary for irritable bowel symptoms. And that's the name of the instrument, and we have one of these measures for each of the 3 subtypes.

This is the foundational qualitative

Page 58

research that I'll be talking about. In our qualitative research, the participants were recruited through GI clinics in 6 U.S. regions and met the following criteria. You can see what they are.

The bottom one, reported an average of abdominal pain intensity score of 3 or more on a 0 to 10 scale over the 7 days before screening. So we did want a symptomatic group and specifically a symptomatic group related to abdominal pain.

One of the first things we did after doing an extensive literature review and interacting with experts in the field, we went out and did concept elicitation interviews with 49 individuals. They were designed to identify relevant signs and symptoms of IBS and determine the way that these signs and symptoms were experienced by patients and how they spoke about them; the relationships between them, the relationships between those signs and symptoms; the most bothersome of the signs and symptoms, the ways in which these signs and symptoms interfere with daily life; and the 5 top

Page 59

signs and symptoms that each participant would want a medication to improve. And you can see the breakdown of participants into the 3 subtypes.

This gives you an indication of what we found. These were the signs and symptoms that were reported by at least 5 individuals, but each of the 49 individuals provided us a list of their top 5 in terms of the signs and symptoms that are most important in their lives to have treated and improved.

You can see that abdominal pain is the first one, and it's universal across the 3 subtypes. The next bar is loose or watery stools, and you can see, as expected, that only IBS-D and IBS-M patients report that as is the case for urgency as well.

We'll talk a little more about these later, but you can see that these are the usual suspects; and again, the types of things that we found in our extensive, as I said, literature review of the research that has already been done, qualitative research with IBS patients.

Page 60

I'm going to only give you a very high level in terms of some very selected findings. One of the goals of this meeting was to talk about the assessment of abdominal pain in IBS, and so I'm focused primarily on abdominal pain.

Across the 3 subtypes, abdominal pain was reported spontaneously by 43 of the 49 participants. Thirty-two of the 49 participants included abdominal pain among the 5 symptoms most important to treat, which is more than any other IBS symptom, and 11 participants identified abdominal pain as their single most bothersome symptom.

In terms of ultimately we needed to then decide, well, what are the signs and symptoms we're going to assess in our measurement tools, in conjunction with our clinical experts, we developed these selection criteria directly attributable to IBS experience and deemed important to treat by most participants within each relevant subtype and that have the potential to respond to treatment within the context of the clinical trial, which is

Page 61

often a 12-week duration.
Note, it was decided that the signs and symptoms included for IBS-M should be a combination of those used for IBS-D and IBS-C.
In terms of the signs and symptoms that were ultimately selected, again, based on the concept elicitation interviews, a review of existing qualitative literature, and clinical expert input, the following signs and symptoms were selected for the draft PRO measures.
They're broken into two areas: abdominal symptoms, pain, discomfort, cramping, and bloating; and then bowel movement-related signs and symptoms, stool frequency, consistency, incomplete bowel movements, urgency, recurrent bowel movements, and straining.
For each subtype, you can see that this is how it broke down in terms of all three of the instruments contained most of the items. IBS-D and IBS-M only have urgency, recurrent bowel movements, and cramping, and then IBS-C and IBS-M are the two tools that contain straining.

Page 62

Note, it's recognized that not all of the signs and symptoms above will be used to derive clinical trial endpoints. Dr. Hanes talked yesterday about the fact that FDA has a concern about urgency, the measurement of urgency, same with straining, but these are symptoms that are important to patients. So we feel that at this point in time -- and again, the final instrument will emerge from the quantitative pilot study.
Our quantitative pilot study will show us how these items are performing psychometrically and how much additional information each of the items is giving us. So there may be some item reduction that occurs. And some of these may go away if, indeed, they're not providing useful information. But we did feel that we needed to go out with this item pool for our quantitative pilot study.
We go from the concepts or the signs and symptoms, and then we have to generate items for each of those signs and symptoms. So multiple alternative items were generated for each of them. The items were then used to assemble the draft PRO

Page 63

measures for further qualitative testing through cognitive interviews, and then the three measures were named, as I said earlier, the diary of irritable bowel syndrome symptoms D, C, and M.
The format and mode of data collection, what we decided upon, was we needed to deploy these on handheld devices. As you will see, or as I mentioned here, the format for entry of bowel movement-related signs and symptoms responses is event driven. So in this case, the event is a bowel movement. So we want them to be able to have a device nearby so that as soon as possible after the event occurs, they can report on the bowel movement-related signs that are part of the instrument.
The format for responding to the abdominal symptoms, pain, discomfort, et cetera, is a 24-hour recall at the end of the day. At that point as well, they would be able to report any bowel movements that they hadn't reported earlier in the day as it had occurred.
We then went out and did cognitive

Page 64

interviews, and so three rounds of cognitive interviews were conducted to confirm the most important signs and symptoms were addressed. We wanted to make sure that we covered what patients felt we needed to be covering and to optimize item wording and response scales.
Some of you are certainly familiar with cognitive interviews, but one of the things we do is we ask people to read aloud the item, and as they're doing it, we ask them to explain to us their thought process as they consider what's being asked of them and what they do, what their process is when they decide what response to give.
We also explored the differences between symptoms, primarily the ones that were talked about yesterday in terms of how are people distinguishing between abdominal pain, abdominal discomfort, abdominal cramping. You can see that we had 43 subjects again broken down by the 3 subtypes.
Again, just some selected findings, although often described as very related, the majority of participants reported a distinction between each of

Page 65

the abdominal symptoms, specifically, the pain, bloating, cramping, and discomfort. For instance, abdominal pain was commonly described as a sharp, tight, or shooting sensation, whereas abdominal discomfort was often described as an irritation, fullness, and/or ache. We have these sorts of distinctions for each of the symptoms that we have included in our instrument.

More selected findings, abdominal pain is a highly salient and important symptom to patients, regardless of IBS subtypes. That certainly was expected. But how do we measure it?

I just want to say I certainly empathized with Dennis when he was talking about herding cats because one of the disadvantages of a consortium approach to the development of a PRO measure is that everyone has a very strong opinion about how each item should be worded.

We have 10 items total across our 3 instruments, 3 measures, and you can't imagine how excruciatingly painful it was for each of those 10 items. And I'm just going to give you an example

Page 66

of this.

During the cognitive interviews, we actually tested 4 different versions of the abdominal pain item, one of them being how would you rate your abdominal pain at its worst in the last 24 hours. We had proponents in the group of using a verbal rating scale as opposed to an NRS, and they just really wanted to see what patients thought about that and whether that might be a better alternative to a 0 to 10 numeric rating scale. That was one of our options.

Then for the numeric rating scale options, the stem was, on average, how would you rate any abdominal pain you experienced in the last 24 hours. And then there were two different essentially sets of descriptors that were used on the extremes of the NRS. The first one was where zero was no abdominal pain and then 10 worst abdominal pain I can imagine, and the other option here was again, zero was no abdominal pain but 10 was worst possible abdominal pain. And then option 4, how would you rate your abdominal pain at its

Page 67

worst in the last 24 hours, and again, this is where we used at the extreme end of 10 worst possible abdominal pain.

Just to give you a sense of how ultimately we decided where we were going to land, one of the things I didn't mention is that our items included initially last 24 hours as opposed to past 24 hours. Again, that was a bone of contention among the group, which should we use.

The words "last" and "past" can be interpreted in different ways. The use of the word "past" most commonly refers to the most recent 24 hours, and so that was confirmed in our cognitive interviews. So the decision was to go with the past 24 hours.

This issue was brought up yesterday as well, on average versus worst. Participants described different methods of averaging their pain over the course of the day. That was one of the concerns, and Dr. Lee Simon brought this up yesterday in terms of in OMERACT, they found that average -- I think you were saying, Lee, that average was what

Page 68

ultimately was landed upon as potentially the best way to go.

We found the exact opposite in the sense that our concern was that, cognitively, people are using all sorts of different ways to decide on what is average, whereas for the most part, we felt that participants were consistently interpreting the word "worst" as their most severe pain during the past 24-hour period. Again, we had a small sample size, 43 individuals, but that was our finding.

Then although participants were generally able to articulate the difference between a symptom at its worst and then on average, they responded the same way or very similarly to both items.

So I think that's important as well, and I think there's a large body of evidence that would indicate that in some respects, it doesn't matter whether you use average or worst because for the most part, you get the same response. So we went with using worst, and that is consistent with what has been the FDA's preference in terms of a 0 to 10 numeric rating scale.

Page 69

Then the whole issue of a numeric rating scale versus a verbal rating scale, so across rounds, there was a slight preference for the NRS, the numeric rating scale, as opposed to the verbal rating scale. But in addition, the NRS is used more often, it's used in clinical practice, and certainly, the FDA IBS guidance used or recommended the NRS. So the NRS was ultimately chosen.

Then this issue of worst abdominal pain I can imagine versus worst possible abdominal pain, although all participants were able to select a response using either version of the numeric rating scale, some participants stated that they could imagine pain more severe than they ever experienced, and thus they would not use the upper end of the scale.

So that's a concern because we certainly want a scale, a response scale for which people will use the full continuum. So the decision was to use worst possible to increase the probability the respondents would use the entire response scale.

Page 70

This was another issue, placement of worst in the item stem, and two participants reported that moving the word "worst" could improve question clarity, and their recommendation was supported by the translators.

For our instruments, we do a translatability assessment. We don't do full translations, but we have translation specialists review the wording of our items and response sets. And in this case, that individual recommended changing the sentence structure to facilitate future translation for multinational trials. So the decision was how would you rate your worst abdominal pain rather than how would you rate your abdominal pain at its worst.

The final item -- and again, this is just the abdominal pain item -- how would you rate your worst abdominal pain in the past 24 hours with the response scale of no abdominal to worst possible abdominal pain.

But we essentially came full circle. This is almost identical to the wording recommended in

Page 71

FDA's IBS guidance, which used an 11-point interest to ask patients to rate their worst abdominal pain over the past 24 hours. The only difference being we used in the past 24 hours as opposed to over the past 24 hours. And this is a general representation of how it shows up on the handheld device.

The limitations of what we've done so far, and again, I've just given you a very high level look at our qualitative research, but although the study participants are reasonably representative of IBS clinical trial population in terms of age, sex, race, ethnicity, and education, 92 people recruited from 6 U.S. clinics are unlikely to fully represent this target population, and we recognize that.

The working group members, again, we were appreciative for the financial support from Allergan, Ironwood, and Takeda, and their representatives that are mentioned here that were very much a part of this process. Then I need to acknowledge the folks at RTI Health Solutions, Sheri Fehnel and Claire Ervin that were a part of

Page 72

this whole process in terms of collecting. The two of them did the interviews, both the cognitive interviews and the concept elicitation interviews, and are now conducting the quantitative pilot study.

Then you can see we have a number of clinicians and other researchers that have helped us with this, as well as many of you probably know Nancy Norton from IFFGD, who was a patient representative on our working group. And Dr. Chey, who is not here right now, was very helpful early on in this process as well.

With that, I will conclude my remarks.

(Applause.)

DR. SMITH: We have a long break for checkout.

(Whereupon, at 9:57 a.m., a recess was taken.)

Q&A and Panel Discussion

DR. SMITH: We're going to get started. I just want to introduce the two members of our panel who haven't spoken yet. Dr. Farrar is an associate

Page 73

professor of epidemiology at the University of Pennsylvania, and Dr. Landis is the professor and director of biostatistics of the Department of Biostatistics, Epidemiology, and Informatics at the University of Pennsylvania as well.

We're just waiting for John.

DR. TURK: It's been a stimulating day and a half. Hopefully, all of you are feeling the same way. Yesterday, there was an orientation to us to think about moving out of our silos to making sure we have a bit more understanding about some of these different conditions that share some common features but in fact are unique in many ways themselves.

We started looking this morning in the presentations at some efforts to tease some things apart in more detail, lessons learned, things we're learning from these different approaches. I think that's been very helpful.

Remember what our objective is. There's going to be a quiz. The objective that you should be thinking about is we want to come up with some

Page 74

type of recommendation, suggestion, ideas about what we want or think would be useful for people to do when it comes to the assessment of outcomes in clinical trials. Actually, it could be other kinds of research as well, but I think clinical trials predominantly in these particular conditions.

So that's what we are trying to get to, and the idea is that we've used the presentations, and more importantly the interactions that people have had over the coffee breaks and over the meals to try to get us to the point where we can move in that direction.

We have moved from a little bit, the silo, and now we're moving back to, okay, how are we going to pull this together in one type of program?

We have a panel. It's nice to see we have a few biostatisticians because we always need them to keep us honest, so thank you for joining us, Dick and John. I think of you as a combination of a biostatistician and a neurologist.

DR. FARRAR: Yeah, I was going to say I think I'm insulting my biostatistics colleagues --

Page 75

DR. TURK: You're the interdisciplinary team all within yourself.

DR. FARRAR: I'm the epidemiologist in this area.

DR. TURK: We want to see from the -- well, first of all, what I want to do is before I ask for your questions out there, comments from anybody on the panel about each other's presentations or anything that you've heard that you think would be wise or useful for us to at least have on the table for discussions and maybe even leading us toward our endpoint.

Dick, you look like you're -- anything you want to say to us, any wisdom for us, comments that you want to make about the presentation -- okay.

John, okay.

DR. FARRAR: The one thing that struck me about both presentations or the presentations this morning is that there actually is a fair amount of information available to think about with regards to what the goal of this particular meeting is, especially with regards to the IBS measures.

Page 76

There's work underway that is going to help inform that process in a very specific and useful method. The work that was done as part of the MAPP spent a lot of time thinking about how measures work and which parts, which measures should be put together into an outcome.

I think that the process of trying to summarize some of what we have heard today may in fact be wait a little bit. There is a process underway to try and help define that.

Then the second thing that was obvious is that the diseases and the processes that we're talking about, even though they all occur in the same general region in the body, are distinct and different. Even within IBS, I think the point has been made very clearly that there are at least two types, and then there's the type that has both, and those are going to be different.

I think that what struck me really was the need to be both general in measures that capture some parts of this, and then more specific for individual components of this.

Page 77

DR. TURK: Thanks, John.
Rick, any comment you want to make?
DR. LANDIS: I really appreciate the opportunity to be here and felt that these presentations this morning really captured the complexity that we're dealing with really well.
What I'm as a statistician very interested in is the fact that these syndromes have multiple domains of symptoms, and the more we try to create a global summary measure without paying attention to the individual target sub-areas of symptoms, I think the more we're missing opportunities to identify the different subtypes of these conditions and the fact that targeted measures for each of the unique subdomains of data are critically important.
One of the things we're discovering in MAPP II that Quentin summarized this morning in his talk is that we have a run-in period with 5 weeks in which there's a screening visit, and then the participants in the next 3 weeks each week log in and do a full battery of symptoms, plus they repeat the body map.

Page 78

So at the fifth week when they come in for the deep phenotyping with all the biomarkers and the QST and the neuroimage scans, we have the background of 5 weekly repeated measures of each of these key features.
I think this will be really useful. We haven't gotten very far because we're still recruiting, but we're beginning to look at the initial one-half of the participants. We now have over 400 who are through the screening visit, and one of the issues is how stable these subtypes are.
When you have a bladder phenotype, is it repeatable, or does it vary from one week to the next? When you have regions on the body map, are they endorsing that same region every week for 5 weeks, or does it rove all over the body? We'll be able to answer those questions now, and I'm really looking forward to that.
But I think that's going to be the key to identifying subtypes that are repeatedly endorsing the same features.
DR. TURK: You want to respond r-- okay.

Page 79

Quentin.
DR. CLEMENS: I'm not responding. I had a separate --
DR. TURK: Oh, okay.
DR. CLEMENS: What struck from Dr. Coons' talk, and we had discussions yesterday, this idea of average pain versus maximum pain or most pain. And I think what you said was it really doesn't matter, but it seems as though the maximum pain is what has been decided upon.
I guess the point would be that if it appears that this issue has come repeatedly in various pain states, perhaps a statement that says, listen, it doesn't really matter, just pick one, and maybe worst pain is a little more understandable.
That would help some of the rest of us, let's say we're going through a similar process for IC or chronic prostatitis, to maybe just have that as the background in a statement from this group or others so we can avoid the perhaps, how do I put it -- shorten the process by a few days by not

Page 80

needing to go through the pain you described.
DR. TURK: Let me just comment that when we developed the draft of the manuscript that comes out of the discussions, that will be circulated to all of you to look at. If for some reason, we've missed any point that anyone feels that you felt got short -- had a second thought about it, or you've got more ideas, there will be an opportunity -- and usually this goes through a couple of iterations. So maybe you'll see it two or three times before this is ready for submission for publication.
So don't feel as if everything that you thought about this and you're flying home on the plane or a week from now when you see a patient -- there will be opportunities to try to bring other things up.
What we will do is there will be a draft manuscript that Jen and Shannon will take the lead on. They've been taking copious notes, minutes of what's going on in the meeting, trying to get this into an initial version. They'll probably

Page 81

circulate it to the steering committee for the first round of comments, and then you'll see it again, so you will have an opportunity.

I'll pull you before I take Dr. Coons' comment, that is, if you look around the room, the number of people are here, that for us to be able to move this manuscript along -- even if you want to say great job, at least let us know you've seen it -- preferably, you'll give us comments on it so we can improve it or clarify things or explain how things are done.

Then there's an attempt to synthesize, harmonize, if you will, the comments to come up with the next version you're going to see, which again you can then look at and then get back to us. The more reasonable turnaround that we have, the more you'll remember your comment and your questions and why you said what you wanted to say. If it ends up taking too long, you're going to forget, or you may forget, some of the concerns you had.

As a plea in advance when you get

Page 82

these -- and make sure we have -- if you change an address or change an email, make sure we know about that because it may be two or three or four months before you see it the next time, but it really means that we need to keep up with you.

I'm sorry, just editorializing that.

Dr. Coons?

DR. COONS: That's okay. I agree totally with Dr. Clemens. It's a situation where this, I would think this issue of average versus worst would be settled science. And from my read -- and it's a superficial read of the literature -- that it appears, first of all, that they -- as long as you're doing it consistently throughout the trial, using average or worst, it's not a problem. But for the most part, the literature that I'm seeing is that they are almost the same score, if not the same score for individuals, when asked at the same time.

I think that an important part of this paper could be just that, that there is a -- I don't know if there's an opportunity to do a more extensive

Page 83

literature review that then there would be an empirical basis for making that statement, that it really doesn't matter which you use. And we'll see if the literature shows that, indeed, people cognitively are coming up with their answer related to average pain very differently, and so maybe we should be concerned about that.

DR. TURK: Shannon, I don't know if this is premature, if you want to even comment, but Shannon Smith has been involved in the process of doing a detailed analysis using the FDA's database to address exactly the issues.

Shannon, do you think it's premature, or do you want to make any comment?

DR. SMITH: It is slightly premature, but I will say what we've done. We did a systematic review of pharmacologic treatments for low back pain, osteoarthritis, fibromyalgia, postherpetic neuralgia, and diabetic peripheral neuropathy -- so from the literature -- that reported both average pain intensity and worse pain intensity.

There are a few people in here who already

Page 84

read the draft. I regret to inform you that it's going to be revised slightly because there were some data issues. So we'll see what that turns up, what in terms of like the greater assay sensitivity of average pain intensity or worse pain intensity.

It is something that we're actually working on right now.

DR. TURK: I misspoke. I said the FDA database, but it was from the published literature, so I apologize for that. There's so many different projects going on that I'm losing track a little bit. But the idea is that we may be able to at least put some data to speak toward that issue based on the analysis that Shannon and the group are working on.

DR. COONS: Right. I think that is an important point, which of them is more sensitive to change within the context of a clinical trial. So if there's empirical evidence there that can help us determine that, then I think that's fantastic.

DR. LANDIS: Just following up on the average versus worst, I noticed the 24-hour period

Page 85

was the reference time frame. I'm wondering if you're asking patients to summarize their previous week whether average and worst would potentially separate.

MALE SPEAKER: No, that's a possibility, and we're not doing that in our work in terms of we're not asking them about their weekly worst or their weekly average.

DR. COONS: There is actually data on that topic. Mark Jensen 10 years ago, I guess, now, maybe longer, did several studies, at least two that I know of, where he asked every day and then asked at the end of the week on average for the week and worst and so on. So there is a published literature. It makes a small difference, but it doesn't make a huge difference.

DR. LANDIS: Even for a whole week?

DR. COONS: Yes. I don't think he went to a month, if anybody knows, but I think a week certainly works. Then there are concerns about memory over a month or longer.

DR. TURK: There are some other studies,

Page 86

too, that have looked at that. We were involved, let's see, a long time ago in which we looked at pain at particular 24-hour period versus up to three months, and we actually showed the relationships were pretty close. They were much better than some people who are into the electronic momentary assessment would lead us to believe they are. So there is a body of literature that addresses that.

Jen, you wanted to comment on --

DR. GEWANDTER: Mine is a different topic, so.

DR. TURK: Is there anything else on this issue? Bob, you were interested in this or you were not?

Lee Simon?

DR. SIMON: One question is not just what the point estimate looks like but what the variability looks like between the two. And in addition, John, I wondered about the variability in the context of how much recall changes that in the context of episodic pain rather than constant pain.

Page 87

A gazillion years ago when I was in Washington, one of the people in my division looked at the question of recall versus 24-hour versus one month and whatever, and it looked in our hands -- all this was done by hand; nothing was electronic in those days. It looked like it recall was a problem, whereas more immediacy of the 24-hour or, at worst, 72 hours was the best evidence that we could get at that time where patients gave consistency with less variability.

It's the variability that worries me more so than the point prevalence.

DR. TURK: John, respond?

MR. FARRAR: No, no.

DR. TURK: We're getting a little into weeds.

DR. FARRAR: The weeds.

DR. TURK: But what it really shows me, if not to all of you, is how complex what we think is a very simple question. Physicians for hundreds of years have asked people to rate your pain a 0 to 10 scale. Rate your pain on a 5-point scale. Is your

Page 88

pain mild or moderate?

We've been thinking that that's a simple question, and the complex -- how many -- 2008 you began working on this, Steve?

DR. COONS: 2009.

DR. TURK: 2009. To see how complex it is, I think is a good reminder to us that when you ask people a subjective response, you get huge range of factors that influence that.

John?

DR. FARRAR: Just one very quick comment, Lee and I talked about this briefly during the break, which is that I think the conclusion of what I've heard at least is that they all work, that different disease processes, whether you have constant up and down variation, as you might have I think with IBS or other syndromes, versus a more constant level of pain might help you decide physiologically which one makes the most sense to look at.

I would argue -- and I don't think we probably want to spend much more time on this. But

Page 89

I would argue that if we understand that they all work and that some decision can be made about which one to use based on the physiology of what you're studying, the combination of biology and measurement science sounds like a good one to me.

DR. TURK: I think we should move from this topic. Obviously, we could spend a lot of time on it.

I think Jen had a comment, and then we'll come to the audience.

DR. GEWANDTER: We can let them go first. That's okay.

DR. TURK: Was yours a comment on anybody else's?

DR. GEWANDTER: We can let them go first.

DR. TURK: She's deferring to you because she wants to get the last word.

(Laughter.)

DR. TURK: Yes?

DR. WIEDERHORN: Roger Wiederhorn, FDA. I spoke with Dr. Landis about this, and he alluded to it in his comments, was the stability of

Page 90

phenotyping. Specifically, with the stability, do patients migrate within and out certain groups only or through all groups in terms of the phenotyping?

Also, this is a short-term study -- well, another question, of course, migrating in and out of phenotyping, is to my knowledge, people don't migrate from no Hunner's ulcers to all Hunner's ulcers once they develop symptoms or vice versa. But that would be an important phenotype migration to document, which I don't believe there's evidence for at this point in time.

Also, there is the interstitial cystitis database, which is a longitudinal prospective cohort, if my epidemiology is correct. You can correct me; I'm probably wrong. But the point is that a lot of patients were studied for up to 10 years.

Do any of these findings help you in terms of the relatively short-term study? I realize they were different criteria and everything, but is there any way you can relate them, glean something from them that would be helpful in terms of symptom

Page 91

stability and subgroup and phenotype stability?

DR. LANDIS: Just continuing a little bit further on this run-in period with the 5 repeated weeks, the painful bladder criteria for filling, the pain increases with filling and the urgency that's the painful urgency component, as well as some of these body map regions, there's quite a bit of variability overall, but there's a subgroup of 40 percent who endorse the same feature every week for 5 weeks in a row. And then there's another 30 percent who 3 out of 5 times endorse the features.

So I think there's variability as a characteristic of a subgroup, and then there's the stability endorsing a every time feature of a sizable subgroup. So another feature could potentially be the persistent presence versus the variable presence that would allow you to identify potentially subgroup differences, but this is all exploratory at this point.

DR. LAI: Roger -- this is Henry Lai. The MAPP study similar to the IC database, is really a

Page 92

treated natural history study. Patients come in and out of treatment within that one year or three years that we're talking about. So you might expect some change because they have multiple things that are changing over time in a phenotype in a classification. That's something important to bear in mind, too.

DR. TURK: Does anybody want to comment about this issue about the phenotype stability? Steve?

DR. BRUEHL: I think this relates to the phenotype stability issue. So if our goal is to identify optimal outcome measures for clinical trials, when you do a clinical trial, you have some entry criteria, I think what I've heard over the last couple of days is that the criteria that are used to determine entry in the studies are not necessarily well-conceived. They may change over time.

If you take that as an issue plus the issue of whether the people meeting those criteria are stable or not and how many overlapping conditions

Page 93

there are, I think that has huge implications at the 10,000-foot level for how we would measure things in trials like this.

Let's say you've got -- pain seems to be common to all of these, so clearly the pain component has to be there. But we've also got the component of some type of disease-specific measure, and maybe it's a urinary urgency. Maybe it's defecation issues. Across conditions, it may differ some, but if these people are moving from condition to condition or have multiple conditions, I guess what I would wonder is whether taking a very broad assessment approach would make sense in order to capture everything that might be informative in the future about what silo they fall into.

Because what if five years from now in the course of doing a study, we refine criteria based on the MAPP study and decide that pelvic pain is this rather than this? Well, now we want to make sure we have information on symptoms to be able to go back and reclassify those diagnostically using

Page 94

those new criteria.

I just wanted to throw that out because I think it relates to this issue of whether there are truly silos or whether these are illusory and overlapping and changeable and what impact that would have on the disease-specific measures you might include.

DR. CLEMENS: I found your comments helpful because focusing on the clinical trial applicability, which is really the main focus of the meeting, which is typically a 6-week to 12-week time period. And while, yes, these phenotypes do change, certainly, if we identify someone with widespread pain, let's say, as an important phenotype, we are not seeing in the short term dramatic fluctuations where someone has widespread pain and a couple weeks later has none.

I think keeping in the context that while, yes, there is some degree of instability, in the context of a 3-month time period, which I think is what we're really talking about, perhaps out to a year with the extended follow-up. But these are,

Page 95

at least from the IC and chronic prostatitis world, pretty stable patients. Even in the ICDB long-term study, only about 20 percent overall actually changed and got better.

I think that that's just useful to keep in mind. These are generally stable chronic patients however we phenotype them. In fact, to some degree, at times we've had to do a fairly substantial amount of effort to be able to identify change or identify a way to look at a variable related to change that won't have everyone being stable in it. So I think this may be useful to keep in mind.

Dick, you can follow up with any comments, but during run-in period, we see some changes, but again, people aren't going from widespread pain to none at all.

DR. LANDIS: Yes. I think it's going to be more how variable they are at the threshold of present or absent. But certainly, it's a challenge to make sure, especially for a clinical trial, that you have the correct baseline phenotype and you

Page 96

have something that captures the level of the primary outcome in a way that when you do -- at primary endpoint, that you can confirm that this is, in fact, a real change or not.

DR. TURK: John?

DR. FARRAR: This conversation reminds me that we need to keep, I think, quite clear and probably separate, although they're related, the difference between defining a phenotype and the variability of the phenotype and then defining the outcome measure.

In pain studies, we study knee pain and hip pain and headache and diabetic neuropathy. We enroll those patients into trials, but the outcome measure is 0 to 10, how much does it hurt measure or BRS or something else.

I would just argue that we are very clear about this need to both have measures that define the phenotypes specifically, but that those definitions of phenotype may have nothing -- will not dictate what the outcome measure necessarily would be or the best outcome measure for that

Page 97

trial.
DR. TURK: Jen?
DR. GEWANDTER: In regards to what Dr. Landis just said, based on the MAPP study and Dr. Coons' experience with interviews, we usually for diabetic neuropathy will do like a week-long run-in, get their average pain on all the says, and if they have a 4, they're in.
Do you think that, based on your experience, you need, A, a longer run-in period for these people, and B, would something that came up yesterday, would it be -- if we want to have a minimum pain severity, would it be on only the days they have any pain or all the days?
DR. LANDIS: John, part of my answer is, picking up on what John just said, classifying the correct phenotype is different than their level of pain. So in particular, we're looking at binary features like do they endorse pain getting worse as the bladder fills or not. That feature is a repeated measure for 5 weeks, but it's not the same as what is their baseline pain for the beginning of

Page 98

the study or their outcome at the end of a study.
So the variability that I'm really concerned about is when you try to stratify patients for a clinical trial and say this is a group that endorses the bladder phenotype, or this is a group that does not endorse the bladder phenotype because, in fact, the therapy may be targeted for the one group relative to the other.
It's that reliability that I'm really talking about when I say the run-in period has opened up some new understanding that there's a group that endorses the bladder phenotype every week, and then there's another group that varies whether or not they believe their pain is getting worse if the bladder fills or not, for example.
DR. GEWANDTER: Right. So I think that maybe my question is a little bit separate then because I think if we're going to make pain one of the outcomes, we need to have a baseline level of pain that's at least moderate in these patients.
I guess the question is if their pain is variable and we only do a week-long run-in to

Page 99

evaluate their pain and they don't make it in the study, are we going to be throwing out a lot of people that we shouldn't be, and should we make the baseline period longer because these conditions are not as necessarily as consistent as, say, diabetic neuropathy?
DR. TURK: Dr. Pontari has been trying to get in for a while.
DR. PONTARI: One of the possible advantages we have with at least prostatitis and IC is that even within the pelvis, on the GUPI, there are 6 or 8 areas. You can get data for location and severity and the pain.
Have there been other pain conditions that looked at -- I don't know -- as opposed to just headache or knee pain, where you've looked at number of sites of pain as being an improvement in addition to the frequency? You can get more information out of that using it as a composite score as opposed to just what's your average pain or what's your worst pain?
DR. TURK: Anyone have an answer?

Page 100

DR. FARRAR: Not specifically, but there have been some studies in acute and chronic pain that have looked at patients' ability to differentiate pain at different sites. If somebody comes in with pain in three different sites, they're able to say my knee pain is better this week, but my headache still hurts.
That's confounded by the fact that if you actually get rid of the knee pain, then the headache might hurt more because it's the only pain. But there is an ability to differentiate.
I think what you're asking, though, is whether looking at the number of sites of pain might be another way of assessing the degree of the abnormality, and I don't know of any studies for that.
DR. TURK: By the way, if I don't call you, it's really hard to see because the lights are so sensitive and the microphones, that you can't use that. So try raising your hand. Yes?
DR. VINCENT: Kate Vincent. I've got two points. The first is about the time scale that

Page 101

we're measuring, and I mentioned this a bit yesterday. About at least 50 percent of our patients are going to be female, and not all of those are going to be on hormonal treatments that will give them a stable hormone state across the month. And we know that IBS, interstitial cystitis, bladder pain syndrome, and any other chronic pelvic pain pathologies often cycle in their symptom severity across the month.

So if we're only going to ask about pain in the last day or pain in the last week, then I think we need some way in which we're controlling for their time in their hormonal cycle to collect those data points. We did a systematic review that we haven't published yet but presented at IASP, showing that about 5 percent of pelvic pain trials, including endometriosis trials, where we should at least be looking at that, actually considered hormonal point and the hormonal cycle in the design of that trial. I just think that's a point we need to be considering.

My second point slightly adds to what you

Page 102

were just saying about pain symptoms. We've talked here all the time about pelvic pain. Actually, to me as a gynecologist, that's a composite of a variety of different symptoms. It's noncyclic pelvic pain. It's dyspareunia, dyschezia, dysmenorrhea, dysuria, and though they may not be part of the definition of IBS -- for example, in my experience, lots of IBS patients will also complain about dyspareunia. But the mechanisms generating those pains might well be different, and they might only respond to certain treatments.

I'm not saying they should be the primary outcomes, but maybe we should be thinking about collecting those as secondary outcomes as well.

DR. TURK: Comment?

DR. CLEMENS: The take-home point, I think, is that we should limit our IC trials to postmenopausal women.

(Laughter.)

DR. VINCENT: Then you have to ask whether they're on HRT or not.

DR. TURK: That was Quentin Clemens.

Page 103

(Laughter.)

DR. TURK: John?

DR. FARRAR: I don't want to stop where we're going, but I did want to make one further comment about something that we've been playing with in the MAPP, which is that you asked about run-in periods. I think that we've found in this observational trial is that a one-week run-in period is probably way too short if you're thinking about what happens to a placebo group treatment because everyone enrolled in the MAPP gets better over the first 4 weeks, everyone, almost without exception. And there isn't any treatment that's -- well, there are ongoing regular treatments, but there's no change in treatment that suddenly happens.

What Quentin presented earlier was that if you ignore that fact, you actually get a different answer to the question of who gets better and who gets worse over time. So I think it raises the question of how long people should be enrolled in gathering data, i.e., getting the love that comes

Page 104

with being in a trial before you actually measure their baseline, and then try and establish a benefit over time. That's an interesting question --

DR. TURK: Does that mean that they're all going to feel better from having attended this meeting? Everybody is going to leave feeling very good because you've entered this project.

DR. LANDIS: I'm feeling better.

(Laughter.)

DR. TURK: It was successful.

DR. LANDIS: In fact, the first MAPP cohort, we didn't have a run-in period, and yet we had biweekly symptom assessment. And the regression to the mean or the feeling better after having just been at the beginning of starting a new trial, or in this case, even an observational study that wasn't a trial, we ended up eliminating the data from baseline week 2, and we used week 4 as the launch period for assessing longitudinal change. So essentially, it's a pseudo run-in period of 4 weeks.

Page 105

In MAPP II, we're seeing the same pattern that the first 4 weeks are basically a stabilizing period where those who start out at higher levels of symptoms are decreasing. There is a group at the low end of the scale, though, who actually gets worse during the run-in period. So it reinforces the fact that probably in these cases I would argue for a four-week run-in period for any clinical trial.

DR. CLEMENS: But, of course, you're only going to lose then from a clinical trial design standpoint because you're not going to be running into people who don't meet -- so if your numeric scale value is 4, let's say you have to be a 4 or more. Well, by definition, you're not going to bring anyone in who's a 1 or 2. So you're going to lose those people who might have worsened.

What's going to happen, all you're going to do is -- in other words, you're not going to have the opportunity to capture those people who started below and worsened. So all you're going to do is lose the people who started at a 4 or 5 and go down

Page 106

to 2.

It's just something that needs to be -- if you're going to do the 4 week, you just have to count on whatever would be a 20 percent attrition rate probably during that time.

DR. TURK: Dr. Dworkin?

DR. DWORKIN: I completely agree that there are all sorts of great reasons to think about a 4-week baseline run-in instead of what we typically do, which is one-week. However, if we're waiting 4 weeks before randomization, patients are going to be really unhappy that they're not getting any treatment, placebo or active, for a month. I think that's a real obstacle that I don't know how to confront.

I think all the reasons everyone has said, regression to the mean, placebo effects, et cetera, is a great reason for a 4-week run-in, but the logistic of doing that is, I think, impractical because the patients are going to say I'm out of here.

DR. TURK: We just tell people that you're

Page 107

getting 4 weeks of placebo while you're waiting for the real treatment. So Ted Kaptchuk would say that might be successful.

DR. CLEMENS: You have the control arm. I don't necessarily view -- I'm not a statistician, but I don't necessarily view the regression to the mean as an issue for a randomized trial where you have a control group, which likely will also demonstrate a regression to the mean, right? So this is more of an important thing for a cohort study. Is that not true?

DR. FARRAR: It's the assay sensitivity. The response to the placebo group has been blamed for failed trials more than anything else, and the response of the placebo group is going to be much -- the MAPP data suggests that most of the response to the placebo group would occur in the first 4 weeks. So as a way of eliminating that complaint about doing clinical trials, a longer run-in.

How to conduct it is an interesting one, and I like Quentin's point, which is that maybe the

Page 108

criteria for getting into the run-in period should be much lower than the criteria for getting into the trial because, in fact, there may be people that get worse over time.

DR. TURK: Michel?

DR. PONTARI: I think what you just asked, though -- what he's saying is that isn't there some placebo effect also in the treatment group that would make those equivalent, correct? So why can't you put -- so in a cohort, yes, we understand that.

So what's the reason in a treatment trial that they don't wash out; that they don't knock each out?

DR. FARRAR: It does. You get the balance in the two groups. It's not going to affect in theory the outcome. You should be able to tell the difference.

The problem is that differentiating between groups depends on where they start, and if the placebo group has a much larger response, then you end up with having more statistical difficulty in looking and finding a difference between treatment

Page 109

and placebo.
DR. DWORKIN: Michel, I'm not sure I believe it, but I think the argument is if the placebo group does so well, your active treatment doesn't have a lot of room to do better. So it's depending on the direction. It's either a floor effect or a ceiling effect.
I don't know that I believe that argument, but John is absolutely right, that that argument has been said thousands of time in the literature as an explanation for a negative clinical trial. It's a kind of the placebo group has done so well because of regression, because of placebo effects, because of natural history, that your drug can't differentiate. That's the argument.
We could have a whole other two-day meeting about whether there's any merit to that one.
DR. TURK: Well, we have nothing else to do. All those who want to stay for two days after this meeting to meet with Bob, we will let you do that.
Quentin?
DR. CLEMENS: I just wanted to bring up the

Page 110

Hunner's lesion patients. The question reminded me of a couple things. The first is that keep in mind that this MAPP study is a one-year study of patients who have already had 8 years of symptoms. To be truly meaningful, we'd need to follow these patients longer, see how they do over a longer period of time, and even MAPP II at 3 years might not be long enough to really answer the questions as well as we want.
There's no question that the Hunner's patients are different. In MAPP I -- but there is some controversy about what exactly a Hunner's lesion is. Some of the sites rely on community physicians to refer patients into this more than others. So for a variety of those reasons, the group decided in MAPP I to not really track or identify or look for whether or not the patients were Hunner's lesion patients.
I think over time as there have been different treatments identified such as cyclosporine for Hunner's lesion patients, we've realized that it's much more important to really

Page 111

identify those. So we are doing that in MAPP II to the ability at least to identify those who we have evidence they have Hunner's lesion patients, understanding that there's going to be a group that we don't know.
I think from a clinical trial standpoint, the important point is that we should definitely identify those as a separate phenotype, whether it's deciding to exclude them or to at least identify them prospectively as a different group and track them differently from the clinical trial because I think the urology world has recognized they are a totally different phenotype, and they may respond totally different to the treatments.
Henry is leading this. I don't know if you have any comments about that.
DR. LAI: I think the MAPP II effort will be really good because the number -- the papers that compare Hunner's lesion to non-Hunner's lesions in terms of the systemic manifestation and that kind of comparisons, really most single center, single investigator, very small number of people with

Page 112

Hunner's lesion, 40, 50 at most. It's very difficult to reach statistical significance of any kind of meaningful comparison.
I think that will be really useful. Our anecdotal experience is that they behave very differently and needs to be treated very differently. The challenge is how to identify them and see if they have a different type of physiology or different phenotype.
DR. TURK: Question in the back. I forgot to say this before. Say your name to make sure that the transcriptionist can get it.
DR. JUGE: Dean Juge from Texas. I wanted to make a point about the run-in periods on the low end and also the high-end patients on the high end. A couple years ago, I was doing studies on topical pain creams, pharmaceutical compounds, and we did patient-reported outcomes. We were using the Brief Pain Inventory, and we were offering it as either a paper copy at the time or they could call in and talk to a nurse. The nurses then were trained in how to take the questions and not lead answers and

Page 113

stuff.
What we found is those that were calling in, where they would have to ask them a question to explain it to them, is that people on the low end, especially the elderly, and that could be 50 and above, tended to under-report until they understood because a lot of times they weren't complainers. So they felt like this is the number I want to get.
Then a group at the very high end who had a pain problem for years tended to run that way because that was the only way as the squeaky wheel that they could get access. But once they're in and seen, after a period of time and they get comfortable, then they got real with what the numbers were to them.
So you're going to see that. That's what we saw in the run-in period is that we started with 2 weeks, and they were constantly on pain meds. Then we sent them pain creams and then started tracking it every 2 weeks for 3 months, and then went monthly after that. And we saw numbers that went negative to what they were saying, and then

Page 114

when they would call up to verify that with the patients, especially on a paper copy, we want to validate this, the patient would say, well, it really wasn't that bad the last time. So they're getting better.
If you threw away the first month and looked at it from second month forward as to how did I do, then you saw some real numbers as opposed to in the beginning.
So we thought about that run-in period or whatever, but we had to keep it the way it was set for the first year we did the data. But you'll see that in the data, and I think that's what you're explaining you're seeing now.
DR. TURK: I take umbrage to saying that people over age 50, having just crossed that threshold, would be in the elderly group.
(Laughter.)
DR. LEMBO: Can I comment on the -- just to go back to the run-in period. So we've done a lot of work in this area because I actually work with Ted. We've been collaborators for about a decade.

Page 115

In one of our studies where we looked at this run-in, accounting for that practitioner-patient relationship, even after six weeks, we still saw continued improvement. So 4 weeks may not be enough without any other intervention. And in the IBS world, it's a lot of the co-interventions that I was talking about yesterday that probably occurs.
The other point is this point about the placebo just washing out is actually not a proven fact, and there is enough evidence now to suggest that it may be other factors that are involved. Not only are there genetic predispositions such as dopamine, which is one of our areas of big interest, where there are clear indicators of who may respond better to a doctor-patient interaction that we're not accounting for. But rarely are these trials truly blinded, and particularly in the GI world.
We talked yesterday about why is it IBS-C or IBS-D that's mainly studied. If you're giving a drug that has some effect on bowel, it's not really

Page 116

a blinded study. We can't really fool ourselves to think that. And once you unblind somebody and you add the placebo effect, you're going to have different results. So the fact that these things are additive has been a major assumption, and we're not actually sure that that's always true.
I would argue for the run-in that we don't actually know. A plain run-in of no intervention of 4 weeks is clearly too long for our IBS patients. We can't take them off drugs for that long. Two weeks is too long.
As I argued yesterday, maybe a placebo run-in might be a better thing to do. We just did this with our rifaximin trial where at baseline, we gave them all placebo. It does affect your results. It does lower the efficacy, and we can't tell if it changed the overall things. But that's something to consider.
I'll leave it at that, but I'm not sure the 4 weeks is appropriate. That's my point.
DR. TURK: Other questions for our panel? Again, the lights are deceiving, so I can't tell

Page 117

whether our voice is carrying and being picked up, so you have to raise your hand in addition to the light going on.

More questions for either this panel about specifically what they -- or even bringing up yesterday to try to again move us forward. Michel?

DR. PONTARI: Has DOOR ever been used in a published trial?

DR. GEWANDTER: Yes, right. I think --

DR. DWORKIN: Yes, I think in antibiotics, infectious disease, not for pain.

DR. GEWANDTER: They're doing AE -- they use it a lot for risk-benefit. That's what they originally developed it for.

DR. JUGE: I just want to make one more comment about when we were doing the review of the patient-reported outcomes and stuff that we had found. We started moving the BPI from paper-based to handheld-based in both platforms for iPhone or for Android.

What we found is that -- I know there's some -- I think OMERACT has some information out

Page 118

there about you almost have to requalify your outcomes reporting when you take a paper-based tool that's been used for years and now throw it out on either the internet, especially an app.

I'll just go to the examples given about how you had to play around with the wording, but also, the information. The BPI asks 4 pain questions, and if your first question is how is your worst pain -- see, on an app, they're going each. They don't sit there at a paper and decide what to read first, and they move forward.

If you ask them their worst pain first -- we're doing this in a testing group to see about moving it forward, and we're working with Academy of Integrated Pain Management, who basically owns the BPI, in trying to qualify it for an app.

If you ask them the worst pain first, then that's their last thought and all the other pain registries come off of that. If you ask them their average first, then they've got a different view to answer the next subsequent question.

Page 119

When you're using an app to ask questions, it's almost like the old test that you had to ask a question four different ways and 80 questions to make sure they're not cheating it. You kind of have to do that with the ones that you're rolling through because you're not letting them go back, and it gets their mind in a certain process.

You can lead your answers on that. It's easy to lead answers to get the positive opinions you want on those apps, too, for some of these studies as it is to get the wrong answer because that frame of mind. If you're looking at paper, you can go up and down a list, but not when you're clicking through and moving forward.

One of the things we played with, especially with past answers, was to throw up on the app, if you're asking for an average versus a past time, give them what their past time was. Instead of them clicking a number, it was a sliding bar. So you gave them their old one, and they slid the bar up or down.

By sliding that bar on that size on the

Page 120

app -- and we're doing like you said, you had to make sure depending on the phones or whatever, the size was right, it recalculated how they felt they were doing, better or worse. So they were saying better or worse by sliding a bar, and we used the temperature bar. So they slid it. It went sideways, not up and down. And we played with up, down, or sideways in apps, and sideways is better.

The temperature bar got better results than asking them to rate it against it, not knowing what they did or asking them to rate a verbiage, not knowing what they did, because they saw where it was last time, oh, am I better than I was last week? Oh, a little bit better, or a lot better.

We didn't tell them what to say. We just said slide the bar to where you feel and gave the two endpoints, and we got different results for that. And I think you're going to see as we move into this computerized age, there's a lot of factors like that that go into doing this, especially the younger crowd that's used to doing apps for everything. They're going to slide that

Page 121

bar different than again, elderly, 55 and above, whatever.
(Laughter.)
DR. JUGE: Whatever range you want to make it, I don't -- we didn't stop at 65, but there's a -- we stopped at the age of people that -- we should have asked a question, and we didn't, how computer literate are you? Do you use Facebook? Do you use your phone? Do you just call with it? Do you do things with it?
People that would do stuff with it would give you different ratings than people that wouldn't. They would all learn to use it, but they would score differently because they're used to those devices. They've got Fitbits. They're tracking everything. They're going to score that slider a lot better. So we expected better results from that group.
DR. TURK: Stephen, from your vast experience of working on these things, how do you respond?
DR. COONS: Well, I think there are a number

Page 122

of issues that you've brought up. One of them, just to say, the FDA wants ultimately all sourced data to be collected electronically, so it's inevitable that we're going to be using electronic data capture devices.
The other issue, there are order effects, you're absolutely right, with questionnaires, but you can have an order effect even on a paper-based questionnaire. But many times, order effects aren't as big of a problem as one might think. But if you're asking about different attributes of pain in a series of questions that only show up one item at a time on a screen-based device, I understand that may be a problem.
There will also be more questionnaires that are developed specifically for electronic data capture, so you're not migrating an existing instrument to an electronic data capture platform. That's why all the instruments we're developing within the PRO Consortium are being developed to be deployed on electronic data capture devices, and there are certain measurement rules that you need

Page 123

to think about as you're developing measures that will only be collected electronically.
I do think that -- and a lot of the studies have shown that older adults, of which I am one, are very savvy, that they don't have necessarily much more of a problem in using electronic data capture devices as younger people.
If you have them sitting in front of a computer and you have them clicking a mouse or something like that, there may be a problem if somebody has Parkinson's. There are things that older adults may have, conditions or diseases that they have that may impact their ability to even use a touchscreen.
I think there are lots of things we need to consider, but there are not insurmountable. This is the future, and we just need to know what the limitations are along the way to getting to the point where we're capturing all of this data electronically.
The fact that so many people have handheld devices -- you're talking about using an app, but

Page 124

I'm assuming that these people went to the app store and downloaded it to their own handheld device?
DR. JUGE: Right.
DR. COONS: That's a very attractive approach in the future as long as you know that to get a representative sample, you may need to deploy devices to people who don't necessarily have a handheld device that can be used with that app.
I think again these are not insurmountable issues, and we're going to get a lot better data because of this issue of -- especially daily diary data that people would fill out the day before they needed to hand it in, even though it was a 24-hour recall, whereas you have date and time stamps on electronic data capture devices so you know exactly when they completed it, and there's better compliance.
DR. TURK: The priming issue is a really fascinating issue. I know there are several questions. But from some of these batteries of questionnaires that you're asking people, imagine

Page 125

that the first questionnaire is about your mood and depression, and your next one is about pain versus the opposite. What's the effect of the priming of having to do that?

I think as we think -- I'll get you, John. As we think of the batteries, the numbers of questionnaires we're asking the people fill out, it's not just the absolute number, but it's also what's the impact of filling out -- in the case, you said the worst pain before you do average pain versus if you ask average versus worst. John?

DR. FARRAR: If I could ask for a specific question, which was the best, worst first or average first? Which gave you the right answer?

(Laughter.)

DR. JUGE: The more consistent answers seemed to come from the average first, but we were just playing with the app. We never got to full development. But average first of a past week -- because it asked for the past, it asked for the last 24 hours, and it asked for now. The BPI asked in multiple modes.

Page 126

So the past was getting them to think about the whole week and getting them away from what their current condition might be, good or bad, and then bringing them to day, the now.

DR. FARRAR: The reason we're asking that is that assuming that we don't use -- what should we call it -- mindwashing or brainwashing to design these apps so that we are leading people to the answer we want, but assuming you don't do that, I guess what I would argue is that one of them might be more consistent than the other and that would be an important thing to know.

Getting back to what we said before, as long as it's consistently used by the same person on the same phone for the entire process, it doesn't really matter if it's slightly different for one person versus another. As long as they both change over time, you have a sense as to whether people are getting better or not.

This argument comes up all the time with the 0 to 10 scale, which I think is a wonderful scale for a clinical trial because it translates across

Page 127

cultures, everybody understands numbers, and it works well. But it's a lousy scale if I want to know whether a patient has a lot of pain after their surgery because I don't know what a 7 is or a 5 is or a 7 or a 10. Is your 7 more than my 5 or not?

The reason that it works is because I'm making the assumption that if you start at 7 and I start at 5 and we both go down with the treatment, then I can say that we both got better. I think we should worry about these things and make sure that we're not misleading patients and giving them a reason to give us the wrong answer. But if we're consistent about it over time, I'm comfortable with the fact that as long as they're using the same method throughout the study, we're likely to get valid answers.

DR. TURK: We're getting into a little bit of the details, but for the last word on this, Bob Dworkin, you want to comment?

Then I think we've heard the complexity

DR. DWORKIN: I have technophobia, so I want

Page 128

to change the subject. Is that okay?

DR. TURK: But just to say we're going to close down this, but I think the point that you've heard a lot of is how complex this is. Every nuance from the wording to the anchors to the order, all can have an effect. John's point is as long as the patient uses it the same way may be less of a concern than looking across patients.

Bob, next, you have a different question?

DR. DWORKIN: I will apologize to Quentin if he showed this data, and I didn't process it. This is a question for Dr. Landis as well.

In the MAPP data, I guess I want to know about three percentages. What is the percentage of these patients who have what could be considered clinically meaningful pain and clinically meaningful urinary abnormalities that concern them? I don't know how we define clinically meaningful. For pain, it might be 3 or greater, and I don't know what it would be for urinary abnormalities.

What is the percentage -- so it's one percentage because they've got both, and then of

Page 129

course, the two other percentages are the percentage of patients that have clinically meaningful abdominal pain but have no urinary abnormalities, and correspondingly, the percentage with clinically meaningful urinary abnormalities but trivial or no pain.

Because it seems to me that those three percentages become important for this afternoon's discussion when we're going to be talking about composite scales like the GUPI versus co-primary endpoints of pain and urination versus complex composite responder analyses like we see in the IBS guidance. Those three percentages, I think, would inform a discussion about what are the optimal endpoints, outcomes in a clinical trial.

I'm sorry if you presented those three percentages.

DR. LANDIS: That's very interesting, especially in these syndromes that have several really correlated but different outcomes. The data that Quentin showed for the functional clusters over one year, the improver group, if you noticed,

Page 130

with the baseline reference of 0 after the run-in period was subtracted, I think the clinically meaningful improvement was clearly there because it was 6 to 8 units of change for that subgroup that was, quote, improver.

But if you look at those who improved on the pain severity and then those who improved on the urinary severity, and you cross-classify those two, only about half of them improved on both at that level. So there's a group that improved on the one but not the other or the other and not the one.

One of the things that I think any clinical trial in this chronic pelvic pain is going to have to deal with is the fact that we're going to need multiple outcomes, and the drug or the therapy may actually target the one and not the other. So the stratification, I think, is also going to be -- this may be an afternoon topic. But it's only about half of them who were in that clinically meaningful change level within the first three months, and they stayed down for the entire year. Half of them tracked that way on both of those

Page 131

outcomes. The other half were one or the other but not both.

DR. TURK: Your numbers are getting pretty small. If I remember, in your improved group, it was like 20 percent of the population or something in that range. Then if you then split that in half, so you're getting pretty thin.

DR. LANDIS: It's interesting because it's about 60 percent in the middle who just vary but neither improve or get worse, and then it's 20 percent in each end that were getting worse and staying worse or getting better and staying better.

DR. TURK: Does that suggest that at baseline, you have these three groups of patients with both and then patients with one or the other?

DR. CLEMENS: I think that the way we could do this, which we haven't yet, is you could define -- so first, you have to define what is a clinically meaningful level of symptoms, and generally, we have numeric rating scales. Usually, the value is 4.

We could propose looking at those with a

Page 132

pain score of 4 or above, those with a urinary score, which we have frequency and urgency. We could look at both, and then those in between.

I think from this discussion standpoint is that would be a surrogate definition for those who would be eligible for a clinical trial, and we would then be able to look at the pain and the urinary phenotype in the degree of overlap. So conceptually, you could set up a trial where they did numeric rating scale of 4 above for pain or urinary and look at that.

I think that's what you're asking. We haven't done that. We have the data, but we haven't done that analysis.

DR. DWORKIN: In a month of you seeing patients, what would you say those three percentages are, the patients in your practice that have clinically meaningful both and the percentage with clinically meaningful pain and no urination, vice versa?

DR. CLEMENS: It varies based on sex, but for the women with IC, the majority are going to be

Page 133

mixed. I would say probably 25, 20 percent would be the urinary only. Virtually all are going to have pain.

I don't know what your thought is, Mike, about that.

DR. PONTARI: Urinary only isn't IC, though. That's OAB and things -- are we talking about -- so if we see someone come in with no pain, we're not considering that this, or do you mean --

DR. CLEMENS: There are philosophical differences. If someone urinates every 20 minutes and they don't have any incontinence, I don't know that that's OAB, but --

DR. PONTARI: No, that's -- what about serious symptoms without pain, isn't that really what we're -- conceptually we consider --

DR. DWORKIN: But their pain is 2 --

DR. PONTARI: Low grade pain, okay.

DR. DWORKIN: That's clinically meaningful.

DR. CLEMENS: But we should do that soon.

DR. TURK: You'll have a lot of data, and you'll be having a lot of fun with these data for a

Page 134

long time.

I was thinking about in the IBS world, has there been any longitudinal study that has that much data that you could begin looking at some of these same things to see if, in fact, one thing, they can learn from the MAPP is not just about your outcomes but the kinds of things that they may want to look back at, at those existing databases. I don't know if we want to go there.

From the IBS world, is there any equivalent kinds of projects there?

DR. LEMBO: Not that I'm aware of, not that follows people for a year without treatment. There's lots of placebo treatment data but --

DR. TURK: That are that extensive evaluations?

DR. LEMBO: Yes, not that extensive, yes.

DR. TURK: Quentin?

DR. CLEMENS: Following up a little bit on the outcome discussion, it seems to me that one of the reasons we're perseverating so much about these numeric rating scales is it's just a single

Page 135

question. The FDA has -- so it's a question for the FDA, and I'm going to be intentionally provocative, so don't get mad at me.

The advantage of a PRO is it has multiple dimensions. It's more than just a single question, and it seems to me that the FDA has highlighted how important developing a PRO is, and then set the bar so high that it's impossible to actually do.

At least within our field, I don't think that a PRO has been developed, and there were comments made during the FDA talks that none of the instruments we use really measure up.

My question is are there examples from other fields, pain fields or otherwise, where they have successfully developed PROs that meet your criteria, and what degree of effort and resources were needed in order to meet that bar?

DR. TURK: Anybody from the FDA want to comment?

DR. WIEDERHORN: Yes. I was involved with the approval of collagenase histolyticum product for Peyronie's disease. We had one endpoint, which

Page 136

was degree of curvature, but we had a PRO that was approved. It was developed. It was an iterative process. It took -- I don't know.

Were you involved in that, sir?

It took four or five years, but we ended up using the Peyronie's disease Bother Scale as one of the endpoints. So we have. You're right. It's extremely difficult, and I know we were involved with MAPP because you had approached about doing a PRO. But I think the problem is it takes a long time, a lot of development. It's not simple.

Kevin Weinfurt and I talked back and forth. He's on the MAPP committee. In fact, I sent him one of Sarrit's slides, the whole approach to this. I think he agreed with us that we -- now, it's not a light undertaking. I think Sarrit showed you this yesterday, because we have to be exact. We have to make sure it's reliable and accurate.

I'm not defending it, but I am saying yes, we have been successful in doing PROs.

DR. COONS: But there have been a lot of drugs approved based on patient-reported outcome

Page 137

measures. I don't want anybody to leave here thinking that that hasn't happened. You think of erectile dysfunction, itching. There are all sorts of pain, obviously. There are all sorts of things that are patient reported that there are no biomarkers for.

The issue is -- and you mentioned they have been approved. Well, they have been accepted as endpoint measures. Qualification is a very different step.

DR. WIEDERHORN: I think again that the problem gets into -- and Dr. Lai alluded to it, is that within IC various gradations, there are a whole bunch of different entities, maybe. That makes it very difficult to establish a PRO because you have to define who you're studying. If it's just like anything else, if it's too broad, you can't focus on it. Peyronie's disease was easy because it's fairly obvious what the disease is.

DR. HERTZ: There have been other situations where PROs and other novel end measures have been developed, and the reason why we have set up

Page 138

the -- or why this whole entire team has developed for these qualifications is because this is not an unusual thing. We have a number of instruments that come in that happened that are novel to the FDA even if not brand-new.

In general, I think when a new instrument is developed, the work that we're asking for is the work that is done to develop a new instrument. There's not something novel about the qualification process that FDA has introduced into the concept.

It's just we have, because of the need, developed guidance and a team of qualified individuals with this kind of background. You don't want me reviewing this. I'm a neurologist. What the heck do I know?

I understand that it's burdensome, and I understand that it's expensive and time consuming. But it's not an FDA process. It's just creating the environment in which we can help to some extent.

Now, I've been involved in some of the meetings for a couple of things that are going on,

Page 139

and I got to say, I was a little surprised in some of them about the questions my colleagues were asking during internal questions. But totally open to hearing about the clinical need, the setting, willing to put it in perspective once they have the information that somebody with the background was able to provide.

I want to push back with the concept that there is something uniquely burdensome about qualification in the context of drug development in the U.S. The good news is once you get there, it just opens it up for use.

Now, some of these programs that are developed and instruments are proprietary. Some of them are open. If you can get to the stage where we've got something that's been adequately qualified, I've been taught, perhaps beaten, into using the word "qualification" over validation, but then getting that work done really does create an opportunity to move forward. And the good news then is everyone has confidence that the instrument is doing what it claims to do.

Page 140

DR. TURK: Jacobs? Kovacs. Sorry.

DR. KOVACS: Sarrit Kovacs, FDA clinical outcome assessments. Drugs are approved based on PRO diaries all the time, and we have approvals on nocturia, for example. Patients are reporting on their nocturnal voids. That's a primary endpoint or co-primary endpoint.

Also, with IBS, CIC, IBS-C, with CSBMs, complete spontaneous bowel movements, abdominal pain. There are approvals. Those are still PROs.

Another example is the Kybella example for submental fat. It was a co-primary with a patient-reported outcome tool as well as a clinician-reported outcome tool looking at the reduction in submental fat, where it was a 2-point grade, I think, improvement that you had to win on both.

So there's some flexibility there. Even if we don't necessarily think that a one-grade improvement is necessarily clinical meaningful, there are some ways where you can use a two-grade improvement, but you could still use the PRO or the CLINRO.

Page 141

DR. TURK: Let me end this session now because we've reached the noontime, and I know for people wanting to check out, this is obviously a prime time. So if you can save your comment and we can start off the noon with your comment. I'm sorry to shut you off, but I just want to make sure that for those who haven't checked out, that you have an opportunity.

I believe that we have now started getting to some ideas about what this manuscript is going to look like, and the fun is over, and then we're going to start herding.

Lunch is back where we had it yesterday. We should be back here promptly at 1:00.

(Whereupon, at 12:00 p.m., a lunch recess was taken.)

Page 142

A F T E R N O O N S E S S I O N

(1:10 p.m.)

Group Discussion

DR. GEWANDTER: If everyone can please take their seats, we're going to get started.

Thank you, everyone, for your participation so far and for coming here today. I think that we've had some really great talks and really productive discussion, and we hope that we're going to be able to make some progress on a consensus for these goals we have here.

What we're hoping that we're going to achieve by the end of this meeting is a consensus on types of primary endpoints we should use in these trials as well as secondary and exploratory endpoints that we think should be included in these trials to try to get some consistency across the trials.

Shannon along with help from Dennis and Bob and I came up with a framework for how to structure the discussion today after listening to what we've heard from you all over the past two days.

Page 143

It seems like vulvodynia is a little bit separate from the other conditions in the challenges, so I think the challenge for the primary endpoints for vulvodynia is what type of provocation might be useful in terms of the primary endpoint and would it be something where we just ask patients about it or experimental, those kinds of questions, versus with the other conditions, although when we asked the speakers to talk in our first -- how we were envisioning the meeting focused mainly on pain, it became very clear throughout the discussion that, obviously, we have to be able to assess these symptoms simultaneously.

So the question is how do we assess pain, and how do we combine that with other symptoms, what are the best methods to do that in order to control type 1 error but still have it be clinically meaningful.

Also, another question that we'd like to address is the time frame of the analysis, considering these conditions are potentially cyclical or have flares. Although generally in

Page 144

other chronic pain conditions, we do a landmark analysis of the last week, is that sufficient for these trials or what should that be?

Then, as I mentioned, secondary endpoints. And then if we have still time, discussion of entry criteria surrounding the endpoint. Just for an example, if we're going to be measuring pain, we need to have a minimum pain intensity or we should all agree that we should have a minimum pain intensity.

That's what we're hoping to cover today. Yes, Lee?

DR. SIMON: I'm just wondering, in your construct that you just created, is it not possible -- and I don't know; I'm not in this field. But is it not possible that you could have a drug for IBS that might be developed that only does pain, only does pain?

DR. GEWANDTER: Yes. Thank you for bringing that up. I think that that's definitely true. You could do that. So if you were going to do pregabalin for IBS, then pain might be your

Page 145

primary, and I think that we could definitely acknowledge that in the paper. But for today's purposes, I think that's kind of -- well, after the discussion this afternoon, probably not straightforward, but compared to everything else we're talking about, maybe a little bit more straightforward.

Of course, we'll acknowledge that in the paper, but I think we want to focus the discussion on for those conditions, if drug affects both or multiple symptoms, how are we going to handle that?

Do you have anything to add?

DR. SMITH: I was just going to say so I think what you're saying is we already kind of agree that if your drug, the mechanism of action is to help treat the pain, pain is the primary endpoint.

DR. SIMON: But it's important to understand, though, that from the creation of a development program to target your pain as the primary outcome, that's great. But you don't have the choice up there in your box of the possibility

Page 146

of having secondary outcomes being all these other things, because we don't know, we really don't know, that if you change pain, you might change other aspects that you would then consider them secondary.

The other question is do you want to protect those secondary outcomes from a labeling point of view to be able to be expressed if, in fact, they're important and they're protected and all the other issues.

I actually think that you've only given two alternatives, methods to combine pain and other symptoms in the context of a primary outcome, but I think that there should be a second box of pain as the primary outcome, and then how you would do all the rest of the stuff. Because that may be important, and you may want to decide to do it in a certain way to protect them to be able to have the FDA consider them important enough to inform and label for.

So we need to be more inclusive than exclusive in the context of structured boxes, I

Page 147

think, up there.

DR. GEWANDTER: I think that's great. Does anyone disagree with that? I think we could put that down as something that we would say is a consensus pretty easily.

DR. DWORKIN: A related question, does the flip of this apply? Is there a box for a drug that improves defecation or urination but has no effect on pain?

DR. SIMON: Absolutely. It should be considered.

DR. GEWANDTER: I think actually I was talking to Dr. Pontari about this at the break, that one way to do this is if you think pain is your most important symptom that your drug is going to affect, you make that your primary, and then you do a gatekeeping type strategy where the next one is a defecation or whatever. And that way, I'm assuming that that means you can put it on the label because you have protected type 1 error. So that would be a strategy in which we could do that.

DR. SMITH: Great. Thank you.

Page 148

DR. HERTZ: I would say that when we're talking about outcomes, it might be safest to discuss what's important and how to structure the study, and not worry quite what goes in labeling because that's probably going to vary depending on standards in the divisions and other factors.

DR. GEWANDTER: That's great. So for the paper, we won't talk about it that way, but I think we could still bring up this concept of doing things hierarchically or identifying -- not just tailoring the outcome to the condition but also what you think the drug is going to affect. I think that we can talk about it in those terms but convey the same information.

Anyone has a question pertaining to this subject?

DR. SIMON: Since I'm not an expert in vulvodynia or IBS, I wonder whether or not the experts could tell us whether or not it is inappropriate to develop a drug that might only deal with pain, or might only deal with dysuria, or might only deal with numbers of defecations, and

Page 149

not have something that covers what we've talked about, which is all of this complex symptomology. This comes up periodically in the kind of work I do, and I wonder whether or not they care about that.

DR. GEWANDTER: You mean like clinically meaningful to patients to do that?

DR. SIMON: No. I think do they want a drug that might only deal with pain, or might only deal with dysuria, or might only deal with the numbers of bowel movements a day as opposed to dealing with the construct of all symptoms and signs that we've talked about that are the domains of measurement that are considered part and parcel to that disease state or syndrome.

DR. GEWANDTER: Yes, Quentin?

DR. CLEMENS: I think the answer for UCPPS is yes, and there are examples that exist already. One would be stakeholder modulation, which is thought to have much more of an impact on urinary frequency than on pain. And we have many, many patients who agreed to undergo that therapy even

Page 150

though we tell them that we're not sure how much of an impact it will have on your pain.

DR. LEMBOW: For IBS, the answer is yes as well. So we have lots of examples of laxatives. Those are drugs that help only bowel habits; antidiarrheals, loperamide, only works on bowel, no effect on pain; and several examples of pain predominant. Antispasmodics mainly affect pain, anecdotally at least. Lyrica has been studied in IBS, has predominant pain effect. So the answer is yes, we'd love a pain drug.

DR. GEWANDTER: Great. Thank you.

DR. SMITH: Is it about this?

DR. BRUEHL: Just another comment on this same issue. So it sounds like in the box up there for IC, UCPPS, IBS, there really would be primary endpoint box 1 pain, box 2 disease-specific symptoms. They could be co-primary, or they could be exclusively one or the other, or they could be sequential.

DR. GEWANDTER: Yes. So we do want to talk a little bit about how to combine symptoms, and I

Page 151

think what you're saying is potentially different methods to do that. But I do want to just take a step back because what we were hoping to do was go to vulvodynia first for the consensus because we've talked so much less about it at this meeting.

Maybe I could open the floor to some of the gynecologists in the room or our people who specialize in vulvodynia to ask what their thoughts are in terms of suggesting things for what good primary endpoints would be for vulvodynia. So we know we have Foster's tampon test. So something like that, how do you think about, what else might be good.

I'm looking at you because you're -- anyone who wants -- or Chris has her hand raised.

MS. VEASLEY: Yes. Chris Veasley. Just to mention that, we did only talk about provoked vulvodynia yesterday, but there really is a need to also develop primary and secondary endpoints for women who have generalized vulvodynia, which I think is going to be a lot easier for this group because they have spontaneous 24-hour pain. And

Page 152

it's not as complicated as having to provoke it. But that population of women with vulvodynia has been largely ignored, both in basic science as well as clinical. And I think there's really a need to do that. I would hate to come away from this process and just do this for provoked vulvodynia and not do it for generalized.

DR. GEWANDTER: Just to clarify that, do you think that there's anything that we could talk about as a group in reference to consistent, all-the-time vulvodynia pain that would be any different from issues that we would talk about -- the things that came up earlier about worst versus average and all these things, anything specific that you would like the group to cover other than acknowledging in the manuscript that this is important --

MS. VEASLEY: And different.

DR. GEWANDTER: -- and different condition and the ways we measure pain now would apply to that?

MS. VEASLEY: I don't think there's

Page 153

anything -- I think it generally mimics some of the other conditions that we've talked about in terms of worst, average, and those types of methods.

DR. GEWANDTER: Perfect. Katy?

DR. VINCENT: That was one of the things that I want to say, maybe not so much about vulvodynia, but I wanted to clarify. Is your chronic pelvic pain syndrome meaning with no associated pathology?

Are we considering things like endometriosis-associated pain where we know the amount of pain is completely disproportionate to the disease we find, and therefore, most of the things we discuss here are just as relevant to that condition?

DR. GEWANDTER: Let me see if I understand what you're saying. Are you saying is are our consensus guidelines only going to focus on the conditions that we spoke about today, or are we hoping that they will be more generalizable to other conditions as well?

DR. VINCENT: Are we thinking about chronic

Page 154

pelvic pain as a symptom, or are we thinking about chronic pelvic pain syndrome where we're saying we've excluded all identifiable types of pathology, which therefore means if you're a woman, you have to have a laparoscopy as part of your entry criteria?

DR. GEWANDTER: My read on what I was hearing yesterday -- and I think this is definitely open for discussion -- is that it would be impossible to exclude all other types of pain because there just wouldn't be any patients, and also, practically, doing a laparoscopy on everybody would maybe not be practical.

I got the feeling that recommending an exclusion criteria based on not being able to have any comorbid pain conditions in the lower abdominal area was not something we wanted to do. Do I have any dissent from that?

DR. TU: Sorry. Can you repeat that again?

DR. GEWANDTER: I got the feeling that from all of our discussions and based on a generalizability and a feasibility standpoint that

Page 155

trying to eliminate all other comorbid pain conditions that could affect the abdominal area would not be something we would recommend in this paper.

DR. VINCENT: I think maybe that's two separate things. I think maybe we're saying if you're doing a study on IBS, you don't want to exclude everyone who's had endometriosis. That's one way of looking at it. The way I was thinking about it is are we actually saying that these recommendations will also apply to trials of endometriosis-associated pain, for example.

DR. GEWANDTER: Yes. Okay. So I think we'd have to ask you guys as the experts. We're coming up with these concepts of how to put two types of symptoms together, and then for vulvodynia, what type of provocation for evoked vulvodynia. If there's place where those recommendations might overlap, we could highlight them in the consensus manuscript, but if there are places where the things that we're seeing are really specific for the conditions we've decided to cover, then they

Page 156

probably wouldn't apply to those areas.

DR. VINCENT: I think studying the populations I see, they don't have clear organ-based symptoms. So lots of my patients will have dysuria, dyschezia, which might be cyclical or might be constant throughout the month. Lots of them will have dyspareunia. So I think that they're just as applicable to any of the chronic pelvic pain syndromes.

DR. GEWANDTER: I think when you say that, one thing that I think about is, well, then what kind of symptoms are you interested in treating and throw it back to what is the mechanism of the drug you're looking at.

So you say a lot of people I see have all these overlapping symptoms. Does that mean you want to do a trial to try to shift on all of these things or -- so I think it kind of depends on the context of the trial that you're doing, how many of the things will apply to any given trial.

DR. VINCENT: Then if we're saying that we're doing trials where the outcome is pain, does

Page 157

it matter what the mechanism of the drug we're looking at is? Because that's going to be affected by all sorts of different drugs.

DR. GEWANDTER: Well, I think -- oh, sorry. Bob, do you want to --

DR. DWORKIN: Katy, I want to make sure I understood. Are you suggesting that there really should be three arrows up there, which is the two we have now, vulvodynia with provoked pain, and then these conditions where there's typically a major component of abnormal urination or defecation. And then there'd be a third arrow to chronic pelvic pain.

We just have, unfortunately, neglected chronic endometriosis-associated pain and maybe some other conditions like that, but that they fit in this article. Is that what you're suggesting?

DR. VINCENT: I want to clarify what I was saying. These recommendations are only for conditions where they have a symptom of pelvic pain but no underlying pathology, or whether we think these recommendations should apply to any trials

Page 158

where pelvic pain is the predominant symptom.

DR. DWORKIN: As an expert, it sounds like you're suggesting they could.

DR. VINCENT: In my view, I think it would be great for the endometriosis world to have some advice from the pain world on how these things should be done.

DR. DWORKIN: So if what we've been talking about for the last two days also applies to those conditions --

DR. VINCENT: In my view, do you agree --

DR. DWORKIN: -- let's add them.

DR. AS-SANIE: I'm sorry. I didn't mean to cut off Frank. Go ahead.

DR. TU: Frank Tu from NorthShore Health. Katy's point's an excellent one, but the list gets longer and longer and longer. It's very problematic. So it's easy to advocate for endometriosis because there are strong patient advocacy groups for it, but adenomyosis, leiomyomas, there are a whole variety of other syndromes.

Page 159

Suzie and I wrote a commentary in the British Journal recently about the fact that you really need a third category of visceral pain when you -- well, certainly a fourth because if there's prostatitis, which is really just male undifferentiated pain somewhere in that hinterland region, bowel and you have bladder. You have to have a uterine category as well, which those four cover everything between men and women.

The easier thing to do, I would argue, would be to follow -- maybe the guys from NeuPSIG can talk about this, something where you're saying we have a probability of what you -- basically say we think we've reasonably excluded other pathology versus we don't think we've excluded reasonable other pathology, which allows you the right size of the trial based on your budget. Because if you can't afford to do ultrasounds and laparoscopies on everyone, but the population is simply too complex to do that on, you actually need to be able to adjust for the fact that you have a certain degree of uncertainty with the data.

Page 160

I don't know exactly how NeuPSIG has adjusted, but they seem to have this interesting idea where they will assign a relative degree of certainty to the diagnosis of neuropathic pain, which I think could be used analogously in this broad CPP category.

DR. RICE: Do you want me to comment from the NeuPSIG or -- it came from that we developed a relatively robust definition, but there are a number of conditions. You can never have a complete certainty about that diagnosis, and there are a number of conditions around that that may or may not be neuropathic depending -- CRPS being the most obvious one.

Because we couldn't really resolve that, that's why the grading system was developed. It was only published a year or so ago, so it'd be interesting to see how much it is actually used for trials and practice.

DR. GEWANDTER: Can I just try to -- do you want to say something, Ursula?

DR. WESSELMANN: I would just leave it

Page 161

really with the pain syndromes that we have and not move on to more complex ones because once these consensus goals are implemented for clinical trials, we can probably learn a lot from it that can then be applied to those pain conditions in the pelvic area that are more complex or require more diagnostic methods really to evaluate them, what exactly it is.

It's kind of like headache because the majority of patients who have headache don't have migraine of headaches. But migraine-type headaches are more easy to diagnose because they have certain characteristics.

So a lot of the research of the clinical trials have focused on those very specific headaches and then the medications that are used. So the treatment approaches that are used are sometimes also implicated for those more diffuse headaches that don't really have a name except for headache.

DR. GEWANDTER: I think maybe we can table this a little bit for now, and we can work it out

Page 162

in the draft. People can make some suggestions. I think we could easily have a statement that says some of these recommendations could easily apply to assessing pain in other pelvic pain conditions without actually getting in detail about how we might apply them.

Yes?

DR. CLEMENS: This will be quick. I just agree. The title of the document is Pelvic Pain, though, and so I think maybe an explicit statement that states that we did not address what might be called gynecologic pelvic pain and maybe list those. Because a gynecologist who reads this is going to say, wait a minute, they've ignored 99 percent of the pelvic pain patients I see in a document that says pelvic pain.

Then there might be opportunity for another potential meeting where we hash out the exclusion criteria for endometriosis and all these others, which is complicated.

DR. SMITH: I think that's a great suggestion. Instead, though, do you think there's

Page 163

a better title that we could use instead of calling it pelvic pain? Would you recommend something else that encompasses -- or should we just say IC, CPPS, IBS, and vulvodynia? I see some agreement with that idea.

DR. TURK: It seemed like that one way to deal with, very interestingly, is to be very clear in your introduction about what this is targeting and acknowledging, as Quentin was saying, that these are these other conditions. Certain circumstances, many of the things we talked about could be relevant, but it was specifically focused on these populations.

Number one, acknowledge it so your gynecologists who look at it don't feel left out, but also imply that some of these may be relevant factors for them to be considering in their studies.

DR. GEWANDTER: Sounds great. Yes?

DR. POLESHUK: This is Ellen Poleshuk. I would also make a plug for acknowledging the discovery you've already made, that there's not

Page 164

enough work that's been done in the area of pelvic pain specifically. You discovered so few trials in your review, and so this would be a good place to point out the need for more work in the area, too.

DR. GEWANDTER: Great. Okay. If there are no more comments on that, maybe we can bring it back to the provoked vulvodynia discussion. You guys want to make some comments?

DR. RAPKIN: The tampon test is a reasonably good provocation method. Obviously, it would be better if you could have intercourse, but so many patients no longer have partners or for various reasons are not able to do that. The adherence and the fact that it has been validated makes it a useful test.

We were just talking about the fact that a certain group of patients with provoked vestibulodynia don't have pain with a tampon, and so that's a fairly small number. Most do, and you said that you got around it by making that an entrance criteria, that they had to have pain with the tampon as opposed to saying, okay, we're going

Page 165

to use a large enough tampon that everyone's going to have pain with this tampon because then your adherence is going to go down, as it would with intercourse.

So I think that, as it's been validated, that would be a reasonable method of provocation.

DR. GEWANDTER: Does anyone have any alternate views or ideas?

(No response.)

DR. GEWANDTER: Okay. I think we could go back then. Maybe we could go to secondary endpoints then in vulvodynia. I think obviously, maybe intercourse in the subset of people who want to be having it would be good, and maybe pain with intercourse, number of times that you have intercourse.

I don't know if there's others that you think -- anyone else thinks we should be collecting for secondaries for vulvodynia. Maybe Chris has an idea? Nat?

DR. KATZ: Sorry, Jen. I just wanted to point out that we seem to have established

Page 166

consensus on the tampon test, and I've also heard that that's a useful test. But none of us have or most of us have not reviewed the performance of that test. So on what basis are we arriving at consensus without actually having reviewed any data on the performance of the test itself?

DR. GEWANDTER: Andrea, do you know how it's been validated?

DR. RAPKIN: There was one paper that was published by Foster's group. I don't remember -- do you remember how many subjects were --

DR. WESSELMANN: I don't know. It's 2009, and it was used also in a clinical trial, but it has been validated.

DR. DWORKIN: I'm an author on that paper, and it was a long time ago. I don't remember a whole lot of details about it.

(Laughter.)

DR. DWORKIN: But to Nat's point, I assume we're kind of considering this as a surrogate endpoint, right, for intercourse? It's hard to

Page 167

imagine that it isn't a surrogate. Sarrit doesn't seem to be here, but if Sarrit was here, I'd ask her what the FDA's criteria are for a surrogate endpoint. Clearly, that would be something that needs to be considered.

I would doubt that -- and I know you're going to say that we shouldn't, but I would doubt that whatever was in our paper back in 2009 is going to satisfy anyone who has a rigorous definition of surrogacy.

DR. KATZ: I'm not disagreeing with the recommendation, maybe as a process. Maybe some information, maybe that paper or some information about the performance of the test could be circulated to the group afterwards just in case anybody has any additional thoughts on it.

DR. GEWANDTER: That sounds like a great idea. Of course, we will always -- if we think that that's not -- that's the best we have right now, but future research in other areas, we could suggest areas for future research if you have some other ideas that you think would be better -- if

Page 168

they were also validated in a certain population would be better, we could also make recommendations for research in those areas as well.

DR. HERTZ: I just want to say not everything is on the same standard as a brand-new PRO. So if you're talking about a surrogate, a surrogate means there's no direct way to assess something, so you need to have something else.

Blood pressure is a -- who cares what a blood pressure is. The problem with blood pressure is that longstanding untreated hypertension results in downstream problems, but we don't make companies with any hypertensive drugs measure downstream problems because we know that measuring the blood pressure serves suitably to anticipate all that.

With something like a tampon test, if that elicits symptoms that you're directly trying to influence, I'm not sure I would even consider it a surrogate. It's a provocative test of a symptom that requires provocation.

If you were going to use that as an outcome for a constellation of symptoms in a syndrome, one

Page 169

would want to know the relatedness of that provocative test to the rest of the syndrome, but if you're targeting that pain, then you're targeting that pain.

Much in the way when we evaluate topical NSAIDs for ankle sprain, we allow the pain to be measured when somebody is standing because that's when they have the pain. I don't consider that a surrogate or a provocative test. That's how they have pain.

I think we need to be very clear on our use of the terms because we don't want to create an undue burden where -- imagine that, FDA doesn't want to create undue burden.

(Laughter.)

DR. HERTZ: But we want to limit the burden to where it's justified.

DR. DWORKIN: I withdraw my use of the word "surrogate."

(Laughter.)

DR. BRUEHL: Quick question that just occurred to me. So are we treating this pain of

Page 170

provocation like allodynia, where it's a yes or no phenomenon, and normal is no and yes is abnormal, or is it something where you'd actually be assessing intensity as an outcome?

DR. WESSELMANN: Intensity.

DR. GEWANDTER: Intensity during the activity. Yes, Rob?

DR. EDWARDS: Sorry. I was just about to ask the same question Steve did. But now that it has been answered, I'll assume we want to be specific about how and with what scale we're measuring the intensity of pain that women in these trials experience with the tampon test. I'm also assuming that if that's a primary endpoint, we'll be setting an entry criterion, an inclusion criterion for the trial on the basis of that.

DR. GEWANDTER: Thank you for bringing that up. We are definitely going to -- well, I hope that everyone will agree that we should have a recommendation that whatever your primary endpoint is going to be, that there should be a minimum severity of that or those symptoms at baseline.

Page 171

If our primary endpoint is going to be the tampon test, then someone has to have a minimum severity while doing the tampon test to get into the trial, which would solve the issue of people who don't have that problem.

As far as standardizing the pain intensity measure, I don't know if that's already been done.

DR. RAPKIN: We're trying to remember whether it's a VRS or an NRS that was used, but it may very well have been a VRS. I think it'd have to be decided.

DR. GEWANDTER: Right. So if it's been -- I don't want to say validated either. Laurie's not here, but I can feel her over my shoulder. Yes, that's an issue, right, like if we're going to suggest the tampon test but we as a group prefer an NRS, I don't know how that work or what we -- Bob, do you have any comments on that?

DR. DWORKIN: My fallible memory is that we used 0 to 10, and then used it in the desipramine lidocaine combination trial. So the tampon test was also used in a 2 by 2 factorial clinical trial.

Page 172

I think it's -- do you remember if it's 0 to 10? Because Ellen was involved in all this, too.

DR. GEWANDTER: I think that brings up a good point. So let's say it was 0 to 10, and in that trial or in the validation study used worst -- or I guess it would be pain right now if it's a tampon test probably, right? So then you don't have to worry about that issue.

DR. TURK: It would seem to me like anything that we recommend that's based on some validated measure, to use the protocol for the assessment, as was the validation, because if they validated on a 0 to 10 scale and we said, no, it should be on a 0 to 5 scale or should be something else, then the validation no longer is applicable.

So whatever we recommend, even with the limitations of it, we have to say it should be performed in whatever the accepted protocol is.

DR. GEWANDTER: It looks like it was done with an NRS.

Frank -- or is it related to this specific thing, Rob?

Page 173

DR. DWORKIN: Yes, but Frank's might be, too.
DR. GEWANDTER: Is yours related to this specific thing, or is this --
DR. TU: Yes, it is.
DR. GEWANDTER: Go ahead.
DR. TU: Frank Tu again from NorthShore. So the 2017 article that's authored by Wesselmann and Pukall is available at Open Access. Why don't we just throw it up on the screen? It's got recommended co-outcome measures and secondary outcomes. I'm looking at the table right now. These are all published from August.
DR. SMITH: Is that something that all of the OB/GYN experts here in the room would agree with? Because if that's the case, why do we need to put it up? We can just reference --
DR. TU: I've seen it for the first time --
DR. SMITH: Oh, I see.
DR. TU: It's already written up by experts. Why don't we start by taking a swing at it? It may be perfectly acceptable.

Page 174

DR. GEWANDTER: Maybe, Ursula, do you have it? Could you give to Valorie? She could put it up on the slide.
FEMALE SPEAKER: I don't think a specific recommendation was made over that.
DR. WESSELMANN: We made a recommendation for the tampon test in there. It was mentioned but not recommended.
DR. GEWANDTER: But there were mentioned outcomes that we could consider and decide if we should recommend them? So maybe like a useful place --
(Crosstalk.)
DR. WESSELMANN: I'll send the paper to Valorie.
DR. DWORKIN: Obviously, this paper that Frank just referred to was distributed. So I guess a reasonable question is, assuming most of us have read the paper, does anyone have any objections to what the recommendations are?
(Crosstalk.)
DR. TU: There's a core table. You can

Page 175

really throw it up on the screen right now. If someone can just get on the internet, I'll send you the link.
DR. GEWANDTER: Send it to Valorie.
(Crosstalk.)
DR. GEWANDTER: Yes, Rob, why don't we talk about yours -- yes.
DR. EDWARDS: One more quick question. At the risk of interfering with the magical consensus building process of day 2 IMMPACT meetings --
MALE SPEAKER: Stifle yourself.
DR. EDWARDS: I probably should, but it's too late now. I'll certainly defer to the real experts in the room and to whatever everyone's recommendation is.
It strikes me that one tricky thing about the tampon test will be likely that the time frame for people's recall of the amount of pain with tampon insertion will differ potentially substantially across women. For some people, they'll be rating pain from that day. Others may be rating their tampon-related pain from several

Page 176

weeks previous.
(Crosstalk.)
DR. EDWARDS: Outstanding.
DR. DWORKIN: Ursula, I'm looking at table 2 now in your article, and the recommendation, unless I'm missing something, isn't for the tampon test. It's for a 0 to 10 scale of vulvovaginal pain during sexual activities in the past month.
DR. WESSELMANN: Right. I'm looking at it right now, too, and they discuss the tampon test a lot. So this paper is slightly different than what we are discussing here in the way that here we want to have a consensus, what might be useful to use as the outcome measures for the FDA, whereas what was written there was also for clinical research. So not for every patient population a tampon test is useful or practical actually to do.
DR. GEWANDTER: When you say not for every population, what do you mean by that? Who wouldn't it be practical for?
DR. SMITH: I think what she's saying that they don't give them a box of tampons and say go

Page 177

home and insert these in a clinical setting or would do that in a clinical trial.
DR. GEWANDTER: Got you. Thank you. I wasn't clear. I think that's a good secondary outcome to have, but I think we all talked about the fact that -- and I'm sure you agree that some people don't have sex. And if they are having pain, they might avoid having sex. So that's not really for a clinical trial probably going to be that great for a primary.
DR. WESSELMANN: I forgot to introduce myself again. Ursula Wesselmann. To measure the pain with sexual intercourse, even if somebody has a sexual partner, is difficult because it's so situationally dependent, and also depending on the lubrication, so you could get potentially very varying results. So a tampon test would be much more standardized.
DR. GEWANDTER: Andrew?
DR. RICE: I know nothing about this topic, but there's something -- there's a little alarm bell just ringing about this tampon test. So one

Page 178

thing I'm personally interested in is the developing world and how these kinds of things translate to other cultures. And I have no idea how this test would translate to a lady living in Afghanistan or South Africa or wherever. It seems very Western orientated is I guess what I'm trying to say.
DR. GEWANDTER: Chris?
MS. VEASLEY: Chris Veasley. The gold standard for assessing provoked vestibulodynia is a cotton swab test, and that's done obviously in a clinical population. I think we'd be remiss not to include that. The idea of doing the tampon test that David Foster included was how are we going to measure this pain in between clinical visits, and that was kind of the best case scenario.
I have two concerns with it. I don't know. I haven't followed the literature as to whether this has been studied since then, but it was only done with one type of tampon, and I'm wondering if it's different between like cardboard applicator and a plastic one, it would be different.

Page 179

The other issue is -- I don't know if David looked at this or not, but one of the therapies in terms of using dilators and other things in women's vulvodynia is the idea of desensitization, that the more you do it, the less fear you have, the less anxiety you have over it, therefore the less pain. And I don't know if you looked at that in the trial or not, or if anyone else has looked at that, but that's certainly an issue to bring up.
DR. GEWANDTER: I think that sounds potentially like we could say recommended primaries would be either the cotton swab test or the tampon test. Do you guys think that -- or do you have a -- no?
DR. RAPKIN: The cotton swab test is really, I think, a surrogate in a way, but it's something that -- there have been some more papers recently suggesting it isn't as well correlated with treatment outcome and improvement and a lot of false positives. Of course, it has been studied more than the tampon test.
I think cotton swab test is a good secondary

Page 180

endpoint. I think it would be nice to have something more similar to the natural situation, either intercourse ideal, but we know that isn't practical or a tampon, and could certainly try to standardize the type of tampon that's used, cardboard or plastic.
DR. GEWANDTER: I think bringing up these considerations, and Shannon and I will take a good look -- or Shannon really is the one who's writing the paper -- will take a good look at the tampon test validation and talk about these considerations. Yes?
DR. WESSELMANN: As far as sensitization is concerned, it could go either way. So it could either be daily dilatation of the vagina or it could be that the vagina is getting more sensitized due to the repeated provocation. But as far as I recall, and that would be something we would have to check, when the test was validated by David Foster, that didn't seem to play a role.
DR. GEWANDTER: As we've talked about, we want to minimize the nonspecific responses in a

Page 181

randomized trial that would be balanced, so hopefully, so maybe not as bad of a thing necessarily.

Valorie, do you have the slide? Oh, you're working on it. Sorry.

While we're waiting for the slide, are there any secondary outcomes other than the things we've talked -- do you want to read what we have?

DR. SMITH: I have intercourse, number of times having intercourse, the cotton swab test as a secondary endpoint. Those are the things I have.

DR. GEWANDTER: I think maybe rating your pain during sex for people who are having --

DR. AS-SANIE: There are standardized measures of sexual function that have to do -- that incorporate arousal, satisfaction, partner relationships, those certainly, I think, I wouldn't consider them primary but secondary could be very useful. Lubrication is part of those measures.

I believe the most widely used is one called the FSFI, and it's been validated, but PROMIS now has multiple measures. All of them are fairly

Page 182

burdensome. The FSFI has, I think, 18 or 20 questions, so they're not super simple, but if you wanted to capture all of the domains using PROMIS, you pretty have to use a similar number of questions.

DR. TU: You want me to read them to you as a -- go through the core outcomes or the secondary --

(Crosstalk.)

DR. RAPKIN: I think while you're waiting, the caveat with the FSFI is not to include a total score because you're dragged down by the fact that if you're not having intercourse when it asks about intercourse pain in the last month.

DR. GEWANDTER: That's great. Thank you. Yes, Katy?

DR. VINCENT: Just while we're doing this as well, we've just looked at the tampon test, and one of the issues that I would have about it -- I don't know how much the validation work has been done -- is that it says no lubrication and a cardboard applicator.

Page 183

So the cardboard applicator is going to be standard, but if the woman is bleeding, for example, she will find it much easier to put in and take out a tampon, whereas if she's not bleeding, she won't and therefore will generate more pain.

I'm not sure that is going to give you a standard response throughout the month.

DR. GEWANDTER: Would a way to address that issue would be to standardize like when in a cycle you enroll people and when they would hit their endpoint, so we could say the caveat would be that be a necessary part of the trial, or is that not sufficient for --

DR. VINCENT: I think you would have to say that you were inserting on a day without bleeding if you wanted to have a valid measure.

DR. GEWANDTER: Great. Thank you for that. That's a great suggestion.

Here we go. So can people see this?

DR. SMITH: Here we go. Hopefully, people can read it. If you have your copy in front of you, that also would be helpful. Pain intensity,

Page 184

pain quality, and affect, so the short form McGill Pain Questionnaire, the VPAQ Pain Descriptor Scale, the 4 VRSs related to pain unpleasantness and distress. Those are the recommended core outcomes. And then pain temporality.

DR. RAPKIN: Core outcome doesn't mean --

DR. GEWANDTER: I guess the question would be -- go ahead.

DR. RAPKIN: Core outcome is not specifically a primary outcome.

DR. GEWANDTER: Right. So I guess the question would be, after looking at this, does anyone have anything else they'd like to nominate for a potential primary? And if not, are there any things on this list that you would say would be like -- you would say they're less of a priority to include as a secondary so we can try to -- I think they're a lot there, so maybe we want to try to be a little bit cognizant of recommending the most important ones.

DR. DWORKIN: If I'm hearing correctly, perhaps the strongest recommendation we could make

Page 185

is that if someone was designing a clinical trial of vulvodynia, we recommend they consider either the tampon test or this measure of provoked intercourse vulvovaginal pain, and that we really don't have an evidence base for recommending this or the tampon test. But we can certainly recommend that these are the two contenders --

DR. GEWANDTER: Sorry. Which one did you say besides the tampon test?

DR. DWORKIN: This item is not tampon test.

DR. GEWANDTER: No. Which one did -- what did you say, the provocation one?

DR. DWORKIN: The first one up here, pain intensity, so that our recommendation would be that someone designing a clinical trial of provoked vestibulodynia should consider either of these two as a primary endpoint.

DR. GEWANDTER: But if you consider this as the primary endpoint, you have to exclude people who aren't having sex.

DR. DWORKIN: The investigator would have to figure that out, exactly.

Page 186

DR. GEWANDTER: Frank?

DR. TU: I would agree with that, but the obvious thing to address Chris' concern about handling generalizable vulvodynia is to have some sort of a simple like it's a -- what do you call like a -- the yes/no sort of like -- you branch your logic out. If the person doesn't have logic that makes sense to evaluate them on the 11-point NRS, you'd go to 2 VPAQ scales and look at worse vulvovaginal and average vulvovaginal pain in a typical month.

This seems to capture all of the things we've talked about in the last couple days. It looks very well done for the pain intensity as a primary endpoint.

DR. GEWANDTER: I think then maybe unless anyone has anything else to bring up, that in terms of primary and secondary vulvodynia, we are good.

Of course, obviously, we're going to make a draft, and we hope as the eminent Dr. Turk mentioned, that you'll all comment on the draft and give us feedback. And if there's anything that we

Page 187

don't cover that you think is really important, you can let us know, and we can try to incorporate that. Ursula?

DR. WESSELMANN: Ursula Wesselmann. Vulvodynia in some ways is different for a clinical trial design than the other two or three pain syndromes in that there are a lot of possibilities actually for topical applications, which is not the case in the others. That's why I think the tampon test might be useful, especially if a topical application is used, and there might be other options if an RO [ph] application is used to measure the primary outcome.

DR. SMITH: Can we go back to the other one? Thank you.

DR. GEWANDTER: Then the only thing I wanted to say because we --

(Crosstalk.)

DR. GEWANDTER: There it is. Also, Nat had put this together for us. Thank you, Nat. And he also had the VQOLs. I think maybe we could add that, too. I don't know if it was up with the

Page 188

other one. I just wanted to check to make sure there weren't any more on there.

Now we want to move back to the more left part of the screen and talk about issues of primary endpoints in the other three conditions. Really, I think it became obvious, as I mentioned before, that as Lee brought up, there will be situations where your drug will target pain, and you want to make that the primary endpoint.

But in situations where either you don't know or you think it might combine both, and you want to have your outcome measure be both, we were thinking for the manuscript that we would summarize the pros and cons of the different methods to do that. So things like using co-primary endpoints; hierarchical gatekeeping; DOOR; using a component composite responder I think is what Laura Lee called it, which is like the IBS guidance, what they recommend right now. Then we would just outline the pros and cons of each because we didn't feel like we really as the group have an evidence base to suggest one over the other.

Page 189

So is that a reasonable thing to do in the consensus manuscript as far as everyone's concerned? Do I have any dissent on that or any comments anyone would like to make?

The only thing I guess I wanted to bring up was a lot of the trials that I reviewed used this composite endpoint, and I think based on what I heard in the past couple days, using a composite where you just make one score out of a bunch of different symptoms or two different symptoms probably wasn't the best way to go.

Does anyone disagree with that statement in that you would want that as one of the options that we think might be recommended?

(No response.)

DR. GEWANDTER: No? Okay.

DR. DWORKIN: That's actually a very strong recommendation, that we're basically saying that total scores like the total score on the GUPI that combines pain and urination, or a total score that would combine pain and defecation abnormalities, we are recommending against.

Page 190

DR. GEWANDTER: Everyone's cool with that?

DR. BUTTERFIELD: I think that's consistent with a lot of what we've heard as well, that there isn't necessarily -- they don't track with each other, and putting them together isn't going to be helpful actually. It's not saying that looking at pain and looking at urinary symptoms are not important. It just means don't put them as a composite.

DR. GEWANDTER: That's great.

DR. DIMITRAKOFF: I would support that statement. I think it's important to keep that as a caveat and probably say that, depending on the findings from the MAPP and the emerging studies, it's important to keep that in mind that the two scores don't --

DR. GEWANDTER: Based on our best evidence right now --

DR. DIMITRAKOFF: -- or best evidence at the time, yes.

DR. TURK: We have a plea from our transcriber to please say your name for everybody.

Page 191

Now, if you've said the same thing several times, very quickly --

DR. LEMBO: This is Tony. I guess the one caveat to this is that, as we heard earlier, not everybody has pain. So it does exclude a significant portion of the population in other diseases.

Now, that would have been the case in IBS, but now with Rome IV, we've made it our entry criteria, made it a requirement to have pain. It actually wouldn't affect IBS, but I just wanted to make sure the other groups didn't feel like it was excluding a large subset of their population.

DR. GEWANDTER: Can we just wait one minute? So that's going to be -- we want to get to this idea of what our entry criteria related to our outcome is going to be. I think that what you're saying is very true, For instance, in IC, if you want to include people who don't have pain and would call it discomfort, then I think the outcome has to be discomfort. It can't be pain, right? You can't put people in the trial who don't have

Page 192

pain and then make pain one of your main outcomes.

So I think that's probably something important for us to discuss, on how we would handle that. But I just want to get to one other thing. Maybe I'm being a little rigid with the boxes. Sorry if I am.

We're going to talk about how we would combine these symptoms, and then I want to talk a little bit about this time frame of analysis thing. Generally, for things like DPN or CIPN -- well, CIPN has nothing for -- I just think of it because it's my thing. But we do a landmark analysis of one week out of 12.

I guess the question for you guys -- obviously, this wouldn't apply for provoked allodynia -- is, is one week enough time when these conditions have a little bit of recurrent pain? Obviously, right now, the way the FDA IBS guidance is they say you want to have a responder on 6 out of 12 weeks because they just don't want to do an endpoint analysis, like a landmark 1-week analysis.

I want to open it up to the floor of what do

Page 193

we think about this, what recommendations we can make, or at least considerations in terms of not making it a landmark of only one week. And maybe Sharon can comment on what she thinks about that.

Sharon?

DR. HERTZ: This area is new in terms of the clinical implications. Obviously, it's different than what we do with landmark analysis in other settings. So I'm actually not going to say anything.

DR. GEWANDTER: Do you guys want to comment on what you think about that in terms of how variable the pain would be or the other symptoms, and if 1-week landmark analysis is sufficient or if we should be thinking of other things?

DR. CLEMENS: If I understand correctly, this is the time frame. During the past week, please rate your symptoms. Is that what you're asking?

DR. GEWANDTER: Or you do a diary over a week, and then you just use that last week in the analysis versus the last, say, 4 weeks maybe, to

Page 194

get a better view of the person's experience.

DR. SMITH: Or the area under the curve where you look at what's happening across the entire time period.

DR. CLEMENS: I guess my feeling would be that we're -- while there's a lot of ongoing analyses, and maybe this change, right now, I think a week time frame has been the standard for IC and prostatitis research. And until there's compelling data to suggest we should do it differently, that would probably be the current suggestion.

DR. GEWANDTER: Mike.

DR. PONTARI: I really like the idea of the area under the curve. Has that been used in other studies, and does it correlate with a JRA, a quality of life? What do we know about that in terms of using that as an endpoint?

DR. GEWANDTER: Bob, do you want to comment? You know a lot more about previous studies than I do.

DR. DWORKIN: I'm sure people have taken an area-under-the-curve approach. My understanding is

Page 195

that the concern about that, for example, for a 3-month trial, would be that you could get a significant difference between treatment and placebo that's driven by, say, the first 4 or 5 weeks and that the difference between treatment and placebo disappears by week 12. And therefore, you've got a treatment that apparently shows efficacy but has no durability.

I think Sharon could comment on this.

(Laughter.)

DR. DWORKIN: So the question is if you look at week 12, you've demonstrated durability, whereas with an area under the curve analysis, there's at least the potential that you don't have durability, but you have a significant difference. But I think Dr. Landis is going to clarify this.

DR. LANDIS: That's actually one of the benefits of the functional clustering profile that Quentin shared this morning, and that is, those who improved in that early phase, in order to be in that group, they had to stay at that improved level the entire rest of the follow-up period. There

Page 196

were other patients in there who went on a lower profile early and then went back up again.

So when you do functional clustering, you capture the level, but you also capture the distance they have to travel at the improved level as well. So if you do something of that order, then basically, you have the amount of improvement but also the persistence of the improvement the whole way to the end.

DR. GEWANDTER: Do you want to comment?

DR. AS-SANIE: This is Suzie As-Sanie. I think, though, regardless of what we decide, I think the paper needs to recognize that this is an incredibly under-studied problem in reproductive age women, because while things like one week have been shown to be sufficient, I think we just simply don't ask.

I think any one of the clinicians here that primarily takes care of women of reproductive age when we ask them clinically, there's huge variability according to where they are in their menstrual cycle. And many women, regardless of

Page 197

whether it's bladder symptoms, GI symptoms, or dysmenorrhea, or chronic daily pelvic pain, their symptoms flare right before and during their menses. And if we don't acknowledge that, we are just missing that problem because we simply haven't asked patients, and then we won't be able to move forward.

I would say that while the evidence that's published might suggest a week is sufficient, clinically, it's probably insufficient and would just at minimum encourage more data collection in women that aren't menstrually suppressed or postmenopausal.

DR. GEWANDTER: Sharon?

DR. HERTZ: That's my point. These are the kinds of things that you need subject matter experts to opine on because the standard that we use for general pain in most of the indications that we get, that last week of 12 weeks is generally okay. But it sounds like here that a reasonable case can be made not just for that we don't know, but that it could really be totally

Page 198

wrong. The generalities are the assessment period and the method of evaluation have to be tailored to the clinical syndrome.

Are 12 weeks enough? That's a standard that's been used and has come under huge criticism for a variety of reasons, but what is a 12-week period in the context of somebody who has cyclic changes? What's the interplay there?

I don't know if there's enough to make a recommendation. It sounds like there's enough to raise the issues for further study.

DR. GEWANDTER: Suzie and then Hanna.

DR. AS-SANIE: Suzie As-Sanie again. And I would just probably go one step further and say that not only should it be tailored to the clinical syndrome, it should be tailored to the population. This should be considered in reproductive age women with any pain condition because when we see these patients, whether or not it's pelvic pain or vulvar pain or chronic abdominal pain, their symptoms often fluctuate. And it's not because it's endometriosis or whatnot. Their symptoms just

Page 199

change over the menstrual cycle.

DR. GEWANDTER: Mike and then Hanna.

DR. PONTARI: It would be helpful if the gynecologists or someone talking about this, could give whatever the best questions to assess that, the best method to make sure you're getting that accurately. I think it would help people who don't do this a lot.

DR. GEWANDTER: Hanna?

DR. GROL-PROKOPCZYK: That's what I was wondering, too. If we don't know enough yet to say start the one week of key measurements 7 days after the period ends, or if we aren't at a point where we can suggest where in the cycle we should be focusing the measurement, then what would you want measured? Would you want just people to keep track of how many days since their last period began?

DR. VINCENT: You can answer that two ways. I think that there's plenty of published data. You can cite Linda LeResche papers, for example, showing that there was a clear cyclicity to lots of different pain symptoms, including fibromyalgia,

Page 200

temporomandibular joint dysfunction. There's an increasing body of literature showing that endogenous hormonal fluctuation and exogenous hormones alter the experience of pain and central processing as well as the symptoms of a clinical pain condition. So we know that there are influences of these factors.

As far as what people's pain does, most chronic pain conditions flare at times of falling or low estrogen, so in the week before the period and as the period starts. But I think if you want to get a full spectrum of what's really going on and what the interaction between hormones and bowel function and hormones and bladder function is, for example, you really have to be collecting a full cycle of data rather than choosing a time that you think is interesting.

DR. DWORKIN: Just to be teeny bit provocative, could we say that for many, if not all of these conditions, what should be considered is if it's a 3-month trial, the endpoint is the last month. So not an area under the curve of 3 months,

Page 201

but last month rather than the last week of a 3-month trial, and that would be mean pain or whatever our measure is.

DR. GROL-PROKOPCZYK: What's your baseline measure then, the first month?

DR. DWORKIN: The baseline measure has to be before patients are randomized, and it's the issue -- in the perfect ideal world, yes, it would be nice to have a baseline of a month. But I think we talked this morning about the practical issues of keeping a patient on no treatment for a month, and that's going to be a struggle.

So realistically, it might be the baseline would be 2 weeks with careful attention to where in the cycle women are. It's going to be --

DR. VINCENT: As long as your outcome is at exactly the same point as your baseline measure and you can time that with days from your last cycle and the length of your last cycle. And at least you've got some form of control for that.

I think ideally and what we do in the trial we're running at the moment is get a weekly rating

Page 202

for the first 4 weeks before they're randomized, and that helps to see who's going to stay in the trial and actually give us the data we want anyway. And patients aren't complaining about it.

DR. WIEDERHORN: Given the argument that also during the first four weeks, you get the inclusion in the trial effect, I would argue that the paper ought to say 4 weeks should be considered and that shorter could be chosen for practical reasons.

I don't think we should obviate the need for it by saying that we think it won't work, because I actually think that in certain circumstances, 4 weeks might work reasonably well. It's just that the issue is trying to figure out how to do it, and I think that's real.

DR. DWORKIN: Four weeks is a pre-randomization baseline. So now you've made a 3-month phase 3 trial 4 months.

DR. WIEDERHORN: Yes.

DR. JUGE: I just want to give another example of following your endpoints. We had a

Page 203

product that was first in its area, if you will, and they came up with a dual endpoint to the study. And how that worked is you're looking at not only early efficacy but does it maintain it over time. So this was 6 months, but you could scale it any way.

At month 3 and month 6, you had to hit 80 percent of that endpoint, and then at month 6, the same thing. So you're really looking at two time points. You got a middle time point. Are they going to meet efficacy, and you have an end. And the people that met that were considered the responder group. So it was fail, not fail. We called it a responder point.

I think it answers a lot of the questions going around here is that that's another option that could be used, but it would give you both the early time point on getting success. If that's a severe pain or whatever, that would be good. But if it's symptoms, they might not only want success, but they want maintenance of that success over time. So it gives you two endpoints instead of

Page 204

that one endpoint there.

So however you span it out, if you want 6 weeks, 8 weeks, 2 months, then you can have two time points. And if there are cyclical involvement, if you did a month period, then you have month 3 and month 6.

So you monitor it through the whole time, but month 3 and 6, you did all of your extensive testing. So they would come in weekly for 4 weeks or whatever it took, but you're getting through whatever their cycle is. You don't have to say you got to start on an off or on day of your cycle. If I'm getting a full month in there, I'll catch that and all that data.

DR. GEWANDTER: Dr. Landis, do you want to comment on that? It looked like you might -- you looked like you wanted to maybe say something. No?

Dr. Landis, no? Did you want to say something? It looked like you did.

DR. LANDIS: No. I think that's consistent with the earlier comment I made about the improver early phase persisting.

Page 205

DR. GEWANDTER: Is your method one that does something similar to what -- I'm sorry; I forget your name -- I think Dean was saying --

DR. JUGE: Dean, yes.

DR. GEWANDTER: -- but more like incorporates the whole time.

DR. JUGE: Well, if you are in the responder or the completer group, if you maintain in both sets --

DR. GEWANDTER: No, I'm asking Dr. Landis.

DR. JUGE: Oh.

DR. GEWANDTER: For instance, depending on the week you pick, it might be different, but if he's looking at using a method that looks at the response and duration over the whole period, that might take a little bit more this whole issue of recurrence and not knowing exactly when the pain is going to be the worst and flares and stuff into account.

DR. LANDIS: It complicates the criteria of it that you could imagine saying in the example you raised about the 3, 6, and 12 months if 12 months

Page 206

happens to be -- or 12 weeks happens to be the primary endpoint, then you would have these intermediate measures where they have to reach criteria and stay below those during those key measurement points.

DR. GEWANDTER: One thing I just wanted to ask -- go ahead.

DR. HERTZ: But that doesn't get to the point that seems to be very specific to a condition that may cycle based on a month's hormonal changes.

So there are two questions here. One is when does it make sense to figure out if something is working and if the effect is sustained for what would be considered a reasonable surrogate for long-term benefit, and then, but how do you measure this particular condition, which is different from low back pain?

What's done in more general settings, acknowledging that chronic pain fluctuates in most conditions; different things will exacerbate it; a lot of those are not well-quantified in clinical studies; and day to day, pain scores vary.

Page 207

So one of the methods applied in this other setting is to take the last week, get daily assessments and average them, and try and cut down some of the noise that way, right?

So that works in that setting. That could completely miss the boat in this setting, so here are some ideas. These are research questions, but some ways to start approaching it to come up with an answer would be to see if there are trends in pain based on where in the cycle a woman is.

Hopefully, for the clinical trialists that are going to be doing these studies, a consistent finding may show up. For instance, the third week of a cycle seems to be traditionally within the worst range, even if it's bad for -- you could identify in the course of a 4-week period if there's consistently one of those weeks that tends to be indicative of worst pain consistently across the population. Even if there may be some individual variability, you could then designate that would be the baseline week, and that would be the efficacy week at the end of the period if the

Page 208

data -- or maybe it's a 2-week period or whatever it is.

Then you would basically enroll subjects and begin their study participation in a synchronized way for that. That could get to reducing some of that variability if there is behavior of the pain that is conducive.

In terms of the durability of effect, you can check multiple times, but you don't have to. You can just pick the distance out, the duration out that you think is adequately predictive of a long-term effect and just do that.

I never recommend just doing the beginning and the end and not doing anything in the middle because of course, that's highly informative. But you don't have to do multiple checks per se unless you have concerns.

So for instance, 3 months is hard enough to keep people in a placebo-controlled study. I won't go into a lot of those issues in this crowd, but we have other conditions that we work on in the division where 2 years is a standard outcome.

Page 209

Well, hello? How easy is that to keep people in study?

So what we'll sometimes do is create a series of assessments, and we'll come up with the shortest period of time we think is reasonable to evaluate efficacy that would both satisfy both some measure of durability of effect and feasibility of keeping your people in the study long enough to not have a major missing data problem.

Then you can, with proper statistical input, use additional longer-term assessments, calculations. So for instance, you can do your primary 3 months, and then if you really want to try and see if the durability makes it to 6 months, you can make that secondary to the 3-month assessment.

So if you lose your population and you lose your power, you're not going to be penalized with a failed study by prioritizing the 6th month. And then you can have a 9-month or a 12, whatever you're interested, and those questions can all be asked. But what you can do is start off with

Page 210

something that is at least conceptually feasible once you've got the other details worked out, and then if you want to have statistical evaluation of the ongoing effect, you can do that in that stepped approach.

DR. GEWANDTER: Is it really quick?

DR. CLEMENS: Real quick. Yes. Maybe this doesn't make sense, but for this cyclical aspect related to menstrual cycle, at a minimum to suggest that subjects when they're measured, that premenopausal women let us know when was their last menstrual cycle, when did it start and the duration, whatever the appropriate variables are.

That might allow for that aspect to be examined or controlled for in a study, and it may not be perfect but may be more feasible than trying to follow someone for a month.

Is that maybe a reasonable suggestion, somewhere in between?

DR. GEWANDTER: Yes. Thank you. That's a very good intro to my summary.

I think what the best going forward going

Page 211

forward will be is that Shannon and I and Bob and Dennis can discuss all of the options that we've talked about, and in the paper, just bring this up as an issue in these set of conditions. It might not be as straightforward as just one baseline week, one endpoint week like we often do in some other conditions.

Then offer some of these alternatives we've talked about as things to consider and things that require future research to validate, and we'll include that in the draft that we send to you. And everyone will have an opportunity, as we keep repeating, to give comments and add things and be constructively critical of and provide feedback on.

I think this is a good place to break for coffee and to use the rest room, and be back at 2:45. Sound good?

DR. TURK: You'll all be invited back in five years when all the things we recommend, all the data come in, and we're going to redo these guidelines.

(Whereupon, at 2:22 p.m., a recess was

Page 212

taken.)

DR. GEWANDTER: That was very good progress. In the interest of keeping it going in time, for secondary endpoints for the three non-vulvodynia conditions, Nat again made this for us. So these are things that could be considered as pain-related secondaries for these conditions. We just wanted to see if anyone objects to recommending any of these.

Then for non-pain-related symptoms, we thought instead of trying to come up with a list here, the experts, urologists and gastroenterologists, could just send us the non-pain secondary endpoints that you would like to see in trials instead of -- yes, Nat?

DR. KATZ: Just in terms of this slide, I just put those up there as random examples. I have no opinion about whether those measures are good or not, or just to provide a framework.

DR. GEWANDTER: Okay. Well, I like a lot of them. Of course, we'll say QST would be based on resources and whatever, or if there's any that

Page 213

people would like to add that they think also should be on this. Nat, yes?

DR. KATZ: Just to provide a little bit more context, the concept here was that, as we were discussing yesterday, all these disorders seem to be specific examples of more general phenomenon, which is some kind of general visceral hypersensitivity. We know that these patients have -- many of them -- most of them have these other more general findings.

So the concept is are there any other patient-reported outcome measures that we should consider in general across these syndromes that might capture the more general phenomenon. Then there's patient-reported outcome measures that could be considered, and then there's sensory testing or evoked pain tests that could be considered.

Then once you're done with that, then you could talk about the specific disorders and what measures might be relevant there. So it's just a framework for discussion.

Page 214

DR. GEWANDTER: Yes. I think that as long as no one objects, I think we could have a section of the paper where we talk about including these secondary endpoints for that purpose of trying to better define this potential phenotype of patients who have more of a central component, and maybe we could add a couple of sentences about how the future might look like where we could potentially move trials to a place where we are doing mechanism-based recruitment and not recruiting based on end-organ disease, and how that might be the future of that area. And by including these things in a lot of trials, we can try to get there even though we're not really there yet.

I think that we can have a section on that because I do think that came up quite a bit in the meeting, and just leaving it out might do a disservice, even though we don't feel like we're all the way there to recommend it as a method now.

Does anyone have any objections to that or any other comments they'd like to add to that?

(No response.)

Page 215

DR. GEWANDTER: No? Okay. Great.

The last thing is this issue of entry criteria, and actually, Shannon and I were talking, and we realized that Lee's point at the very beginning, we were remiss in making this potentially a symmetrical diagram, that it's not just places where drugs of mechanisms that we think pain should be the primary, but also potentially defecation only or urination only might be also a situation where you wouldn't expect your drug to help pain but you would expect it to help these other symptoms.

We proposed to put in the manuscript to make our modified figure symmetrical and add how that sometimes you might consider those endpoints only in a trial. Do you guys as experts disagree with that?

Yes, you're shaking your head yes. Can you comment on why you might?

DR. PONTARI: I'm not sure that we would do an IC trial or prostatitis trial just for the urinary symptoms. I know these people have pain;

Page 216

we talked about that. But there's been so many other drugs studied for urinary symptoms, for frequency, urgency, things like that, that I'm not sure -- you can disagree if you want, Henry, but I don't think we'd ever set out for a drug just for urination.

Now, there was a drug that was tested for prostatitis that was an alpha blocker that helped urination. It also helped pain.

DR. LAI: I agree with you. I think it's purely urinary symptoms. It shouldn't be IC or CP. The question becomes the discomfort part and the pressure part. You say pain, pressure, discomfort, plus urinary symptoms, I think it's okay.

DR. GEWANDTER: Okay. So that leads us to our next topic. I think I've said a couple of times how I think if your pain is going to be an outcome, you need to have a minimum pain severity in your trial. I think what Dr. Lembo brought up a little while ago, that I tabled now, is this issue of but we'll be excluding a lot of patients and what do we with that.

Page 217

That's why we were thinking maybe you would only have your outcome be urination or defecation. Maybe that would take care of that group. But it seems that another way to handle that would be instead of making pain an outcome, a co-primary or however you decide to do hierarchical, whatever, discomfort or something else as the primary.

I think we don't necessarily know how to measure discomfort yet. I think yesterday it seemed like everyone was in agreement that that is still elusive. Maybe we could have a section of the manuscript that says something -- oh, Sharon, why don't you go ahead.

DR. HERTZ: I'm not sure I understand how this would work. If you have a population with a condition, and there's these different subpopulations, and some have pain and others don't, you have a drug that's targeting pain, why would you include people without pain? If you did, how could you possibly hope to win? If you have people with discomfort and people with pain, and you have something that targets everything, then

Page 218

you would come up with a discomfort scale.

Now, if your patients with pain don't acknowledge the presence of discomfort, the populations are not necessarily amenable or are they, and if so, how, to being included in the same clinical study. At the end of the day, you need a primary endpoint that is appropriate for your whole population even if you're going to do some subpopulation analyses.

So I'm not sure how you can solve this problem without somebody describing what is the right outcome that the whole population can be assessed on.

DR. GEWANDTER: This is the CIPN problem I have, right, like the same issue. Some people have pain, some people have numbness, some people have tingling, overlap, and how do we handle that, right?

I think what you're saying is are they distinct people, could they be the same people, and could we somehow come up with a single primary endpoint that would incorporate all of them?

Page 219

Yes?

DR. PONTARI: I think from what we had talked -- so we say urinary -- all these people for us have pain. What the category -- what Dr. Dworkin and I talked about were people have low pain. I kind of agree that, thinking about it, there may be people with, let's say, a pain score of 2 with a lot of urinary symptoms. They're not getting into trials is what you're saying. That's like the pain group.

Another thing is what's interesting for us -- and we can comment -- we don't distinguish pain and discomfort. Should we be doing that? Do we have -- our symptoms scores, it's all pain and discomfort. We have no just discomfort and just pain. Is that something that we need to –

DR. GEWANDTER: Change that.

DR. HERTZ: In people who are coming in with a pain score of 1 or 2, they're just going to kill your study. You're never going to show efficacy if that's your primary because they're not getting enough movement.

Page 220

DR. PONTARI: Right, we're not going to do pain in them. I think you were talking about there could be patients with --

DR. GEWANDTER: What I was bringing up was this issue that has come up a couple times that if we -- that some people describe it as discomfort and not pain, and that we're not aiming to create drugs for those people if we just ignore them. They would be excluded in all of the trials if our endpoint is pain.

I think the question is do we as a group -- maybe the question is -- maybe the answer is we know. We just ignore it, and we don't do that, like you have to have pain to get in the trials we're talking about. Or do we want to have a section about how future studies -- looking at how to measure this lower level, something that patients don't describe as pain but is discomfort?

DR. DIMITRAKOFF: I think part of the discussion yesterday was that we simply don't have a way of measuring discomfort at this time. So I don't think we can just say these people should be

Page 221

excluded unless we think discomfort is a different degree of pain.

DR. GEWANDTER: I think that might be true.

DR. HERTZ: Right. I'm not saying that people should be discarded, and I'm not saying that it's not important to consider how to develop therapeutics for them. But at the end of the day in the context of a clinical study, you have to have a primary endpoint, and you have to have people who come into the study with enough of something that can then be changed over time so that you can demonstrate a difference from placebo or whatever your control is.

Given everything that's been said about the placebo effect, regression to mean, and everything else, if you allow people who have on a 10-point scale 1 and 2 symptom ratings in, and that's your primary, you might as well give up because the power to show a change is going to be -- you're going to need thousands of patients.

What is the priority then?

DR. GEWANDTER: So maybe I opened a can of

Page 222

worms that was totally unnecessary by bringing this up. Hanna?

DR. GROL-PROKOPCZYK: Hanna Grol-Prokopczyk. From what I've heard in the last two days, it seems like it's not necessarily clear when discomfort means that someone has low level pain that they're too stoical to refer to as pain. It's really along the same unidimensional scale, but they prefer a different word and when it's measuring something qualitatively different.

If I'm right that that's not always clear, then it seems the best our group could do is mention that that might be an area for future research.

DR. GEWANDTER: I think that sounds like a great plan. I think we're going to be saying that the focus of these research studies that we are talking about is pain, so you need to have a minimum entry of pain to get in the study, and that future research could look into these other symptoms that are similar to pain but may be a little different if that's a different population.

Page 223

DR. HERTZ: Or specifically perhaps, people who have low levels of pain along with the other symptoms, that the research agenda include how does one study them.

DR. GEWANDTER: Actually, following up with that, just to be clear because you guys -- actually, you didn't comment on the issue of would you want to design a drug only for defecation and not -- you guys said you wouldn't want to be focusing on drugs only for urination. So you think you don't want that to be in the paper at all, that concept?

No? Well, you can comment later.

DR. TU: This is a pain meeting. Is that not implicit, what we're doing here? Sorry. This is Frank Tu.

DR. GEWANDTER: Let's save that for the paper, and you guys can comment.

(Laughter.)

DR. GEWANDTER: Sorry. Ian, you were going to say something. Moving on.

DR. GILRON: I was just trying to suggest

Page 224

that maybe there should be a caveat in the paper that could say if someone has a therapeutic agent that the mechanism is likely to address only one of multiple symptoms, that that should be encouraged if that's a possibility, and that will affect the trial design. We're not necessarily looking -- people may not necessarily only have agents that are going to cover the whole spectrum.

DR. GEWANDTER: Yes, we could be a little more general with that explanation and then use pain as a good example of that. Sounds great.

Yes, Stephen?

DR. COONS: This is Stephen Coons from the Critical Path Institute. But we still need to assess the other symptoms.

DR. GEWANDTER: Yes, of course.

DR. SMITH: To the degree we can. Discomfort is still going to be one of those things that there's going to be a research agenda.

DR. GEWANDTER: I think he means like urination.

DR. SMITH: Oh, right. You mean other

Page 225

symptoms specific --
DR. COONS: So we can prove that we haven't made anything else worse.
DR. SMITH: Oh, yes.
DR. GEWANDTER: Ursula?
DR. WESSELMANN: I was going to say the same thing. We focus on pain, but some patients have pain and discomfort. So it's really two different things and not discomfort being the lower level pain, so it will be important to measure that as well. I don't think that has been really done systematically.
I forgot to introduce myself, Ursula Wesselmann.
DR. GEWANDTER: I think people can recognize your accent.
How to measure discomfort is an area of future research, I think we can all agree on that. Any dissent?
(No response.)
DR. GEWANDTER: No? Okay.
DR. SMITH: I think really the last thing is

Page 226

to ask if there are other ideas that people have for research agendas relevant to the things that we've been talking about here today. We were talking about this need to figure out discomfort, bloating, cramping, and using some of the outcomes that were on the slide that Nat had made. A lot of that would probably be very exploratory as well.
Other thoughts about things that we should put in? Again, you'll get to see the manuscript a number of times, and you'd be able to provide your input along the way. But if there are thoughts you have now about things that we might want to consider for a research agenda as we're crafting the manuscript, that would be helpful. Ian?
DR. GILRON: Ian Gilron. I just wonder whether -- there are a lot of issues that we can learn from previous IMMPACT and ACTTION recommendations and meetings. One of the biggest concerns that comes to my mind with multiple outcomes is the question of assay sensitivity that we really worry about.
I wonder whether we can revisit our table 1

Page 227

from our 2012 assay sensitivity paper and see whether any of those levels of evidence -- we had various levels of evidence for different recommendations like extremes of pain on entry; see whether there's any new evidence to upgrade or downgrade those, and also try to see how they're relevant to these multi-symptom conditions.
DR. GEWANDTER: I think that's a really good suggestion. I think Shannon and I can go through the table together and see if any seem particularly relevant to this condition, and we can put that in the paper if we find things, and you guys can all comment on that, too. I think that's a great suggestion. Thank you.
Yes, Tara.
DR. ALTEPETER: I wanted to come back to a comment that was made yesterday when someone had asked if there were creative ways in which you can assess if multiple people in the trial have more than one symptom that's most important to them. I didn't get a chance to comment at the time, but we have seen really creative strategies to approach

Page 228

that.
I wonder when people were asking the question about how could you broaden your enrollment population to be more representative of the ultimate patient population and wanting to include some of these people who have only low level pain but may be more bothered by their other symptom, I think that if you truly had a drug and you understood the biologic mechanism of the disease that you're talking about, and you know that your drug has a reasonable chance of affecting both of these things, then I think it is possible to potentially broaden your population to maybe some of those people who have less severe pain, but their success or failure is going to be assessed based on what they identified as the most bothersome symptom.
You could potentially have a more heterogeneous population that's enrolled and then say, okay, everybody is going to decide at the beginning which of these symptoms is most bothersome to you, and that would be the way that

Page 229

your individual responder status would be determined.

We have seen at least some proposals for strategies like that. I think it's something at least people could consider if you want to maybe think about how you could get at the idea of having a more representative sample rather than trying to be really homogenous and just take the most severe part of your population.

DR. HERTZ: What is your conclusion at the end of a study, that the drug treats the syndrome regard -- I'm just wondering how one would interpret that outcome if it affects pain in some, urinary symptoms in another, some other distant pain in another, but not -- I'm having a hard time wrapping my head around it.

DR. ALTEPETER: I think it would probably be most appropriate to a symptomatic condition, and you could say that patients had a reduction in their most bothersome symptoms.

DR. GEWANDTER: Yes?

DR. BROWN: Yes. This is Cole,

Page 230

Philadelphia. Just to add on to what she was saying, I think in evaluation of migraine, you see where you evaluate to pain freedom, and then you can also look at the most bothersome symptom, whether that be nausea, vomiting, photophobia. And you're really thinking of it from a syndrome sense, right, as a complex. We've seen that in the migraine space.

DR. GEWANDTER: Is that part of the primary in the migraine space?

DR. BROWN: Yes. So you can have several -- you'll have co-primaries.

DR. HERTZ: Co-primaries are an end phenomenon, so that doesn't quite get at what happens when some people don't have some of the symptoms. There's way to handle multiple aspects of a syndrome and create this paradigm where you have to really get at the critical aspects of the syndrome.

The part here that I'm having trouble picturing is when the manifestations are sufficiently different so that different people

Page 231

actually have non-overlapping symptoms.

DR. ALTEPETER: I guess I wasn't trying to say that they were not overlapping. I was trying to say that if you have some who are much more bothered by symptom A versus symptom B, that you could primarily assess based on improvement in the part that was most bothersome to them.

DR. HERTZ: Isn't that still non-overlapping? If 10 people are being assessed for the urinary frequency and that's their most bothersome, and 10 people are being assessed for pain because that's their most bothersome, if these people don't have a change in their urinary frequency and these people don't have a change in their mild pain, what am I actually measuring at the end of the study? What is the drug doing?

DR. ALTEPETER: I would envision that you could say that the drug is improving the aspect of the syndrome that is most bothersome to them. You would have to believe that your drug has the biological effect on both, and it's just that for your people who are primarily bothered by frequency

Page 232

and have low level of pain, you're not able to detect much change there because it was already at a minimal level where the measurement problem exists there.

DR. GEWANDTER: Maybe Dr. Landis can comment on this and Dr. Coons.

DR. LANDIS: I don't want to complicate the answer to your question by saying the subgroup phenotyping at the beginning of this topic has to also be done in a precise enough way that we have subgroups that are enriched for higher probability of success on a particular drug.

I'm thinking back more than 10 years ago to the IC trial where it was a combination trial for both hydroxyzine and Flomax. We had a primary outcome that failed, but if we now take everything we learned in recent years, you would want to have primary endpoints for those who are going to get better on pain different from those who are going to get better on their urinary symptoms. But you also have to know who those patients are at baseline and stratify them in a way that you enrich

Page 233

so that you don't have the pain subgroup only on the drug that doesn't even target pain.
I think this is going to require the subtyping and stratification at baseline before randomization so that you're enriching for the outcome for the drug class that's being tested.
DR. LAI: Henry Lai. I think we know as clinical experience, there are treatments that are being commonly utilized to treat IC that doesn't improve pain but improves the urinary symptoms a lot. People will go for that, and it's done routinely. These people come in with pain and urinary symptoms, but the pain doesn't get any better with that treatment.
DR. HERTZ: So your primary are the urinary symptoms. That seems pretty clear.
DR. LANDIS: We would need a primary for each.
DR. LAI: You would need a primary for each but not in the N sense and not in the composite way because you would wash out anything that you would detect, because there are mechanisms like

Page 234

neuromodulation that will improve frequency and urgency tremendously without doing much for pain.
DR. GEWANDTER: I think Dr. Coons wanted to say something, too, about this.
DR. COONS: Stephen Coons, Critical Path Institute. I just wanted to follow up on what Cole had.
Sharon, was the migraine guidance out of your group?
DR. HERTZ: Yes.
DR. COONS: Okay. Because that is an important document in the sense that this is a situation where headache is a given. It's essentially combining headache and then the most bothersome of three symptoms: photophobia, phonophobia, and nausea. And the patient at the beginning of the trial would pick one of those three because they're likely to have one of them that is the most predominant and bothersome to them. That is what the guidance recommends.
The only other thing, Osphena was another drug, which is for painful intercourse

Page 235

post-menopause. So there was a biomarker, but then there was also a situation where the patient picked one of three symptoms that was most bothersome to them. So it has been done.
DR. HERTZ: Right, but it sounds like the primary has one common element and then the other manifestations in addition. What I'm hearing is, conceptually, there might be two different primaries based on that prespecification, and I guess the devil's in the details of how one would structure that kind of clinical study.
DR. COONS: I assumed it ended up being a composite endpoint --
DR. GEWANDTER: Maybe we can look at the headache guidance and look at the details and see what kind of example we can get from that.
Also, maybe, Tara, if you could send us some of the examples. I don't know if they're proprietary things, but if there's something that you could send us that we could look at the details, maybe we can try to incorporate some example like that in the paper after talking about

Page 236

it with our steering committee and then everyone can comment. I think that would be a good way to move forward with that subject.
DR. BUTTERFIELD: Just a quick comment. It's Noam Butterfield. It's actually in the pre-read that we got for this meeting. It's in the multiple endpoints document where they give the example of the migraines.
DR. GEWANDTER: Migraine, great.
DR. BUTTERFIELD: Line 521.
DR. GEWANDTER: Thank you.
DR. BUTTERFIELD: So you can get it quick. I think the point was just not necessarily -- it may not be that we're looking at two different primary endpoints, just are there methods to look at a primary endpoint and additional ways to look at those additional symptoms.
One way to do it may be not just choosing one because maybe that one urinary symptom, for example, is not the same or not the most bothersome to all patients. So some patients maybe it's nocturia; other patients, it's frequency. Maybe

Page 237

what's the most bothersome to that particular patient is one of those secondary variables, and ranking them by most bothersome is a method of doing it.

DR. GEWANDTER: Yes. You hit on something that we intentionally glossed over. We didn't even get into how we should recommend measuring urination abnormalities and defecation abnormalities. But I think because of the focus of our group and this meeting, that we're going to leave that as not -- we're not going to define that in this paper. But I think that your point is well taken that maybe that can be a helpful way to do that half of the symptoms as well.

Does anyone have anything else they'd like to bring up? Yes?

DR. TU: Can I bring up one last thought related to some of these measures? Frank Tu again from NorthShore. There was a presentation by Bill Chey, who I don't see here unfortunately today, about some rather interesting app where you can grab a lot of these secondary measures that was for

Page 238

GI specifically.

Quentin didn't mention this, but MAPP has an app as well that grabs urological measures. There's actually two different forms of that being used on MAPP. There's a group out of Medical College of Wisconsin that's built another symptom tracker.

Is it within the scope of this recommendation to talk about the idea of trying to get more patient-facing data collection where a patient could do it on their own and come into trial having already phenotyped themselves as a next generation strategy trying to minimize cost burden of these trials?

One of the problems of this group of people, as we've talked about a lot, is that there's too many things packed into the pelvis and abdomen. One solution that we might propose is that future groups need to just fundamentally change the game and have the patients get the data themselves and essentially free themselves up from the research teams.

Page 239

DR. GEWANDTER: Are you saying, in essence, patients fill out PROs on an app before they come to their first visit?

DR. TU: They can do it through their whole life if they're really that -- especially like the ones that we talked about, the true comorbid conditions, there might be a call to action that we could put as part of this, to say it's so complicated to study this group of patients that one potential novel avenue that we would propose needs to be an area of significant inquiry is how to create an infrastructure that severs patients from clinical research in order to track their own symptoms and to make that in some sort of universal code that can be pulled into trials subsequently.

DR. GEWANDTER: I think that maybe what you're saying is it's kind of like -- what is that term? They do it a lot in other countries where they have an infrastructure set up where it's a registry trial. They already have a registry, and then they randomize within it.

I think that might be a little outside the

Page 240

scope of this paper, but I think it's a very interesting point. But I think that's something we'd have to think about, how it might fit in the paper. But I think it's well taken, especially if we're thinking about baseline of 4 weeks, maybe that would fit in there.

DR. DWORKIN: One of the things we're going to have to say in this paper is that we're focusing on outcomes and that there are all sorts of research design questions that were beyond the scope of this effort, but that could be the focus of a subsequent effort.

Yesterday, I think Sharon mentioned enriched enrollment randomized withdrawal trials. That's the kind of thing we didn't talk about at all at this meeting, but would be worth considering at another meeting.

DR. GEWANDTER: All right. Well, thanks, everyone.

Oh, sorry. One more thing.

DR. JUGE: I just wanted to make one more comment, and it's about the PROs. It was a comment

Page 241

I was going to make before we broke for lunch, so I'm sorry for keeping you guys now because we're about to leave.

But the PRO issue I think changed in around 2008-2009 when the FDA allowed that to be part of the indications. Because what happened is you got a drug approved, and then you did the outcome studies and added data like that afterward. But it took a couple of years to get the data. And I believe it was in 2008 they allowed the combination.

So if you're doing that at the same time as you're doing your phase 3 indication, you can include that info into the label. So companies started looking at putting that info in the label, and they started designing these PROs for studies. But once the study is done, what do you use out in the field? If it's so cumbersome, nobody is going to touch it.

From my perspective, the data that we did is we took the useful tool that could be used in the field and build it backward, and see how I can wrap

Page 242

it up for a study but make it a useful tool. So my pitch is if we're going to make a statement about coming up with guidelines, then we should also develop a tool that can be used forward because manufacturers want two things out of the PRO. They want the information in the label from an outcomes basis because that's where everything is going, what's helpful to the patient.

They also would love a tool that would then allow them to fight with managed care plans that are doing prior auths to say I have a tool that if I'm doing that will show the benefit of this drug. So when you come to me in a year and want to approve for the refill, I can give you that data to continue the use of that versus patients not getting benefit, I should drop them.

That stuff will help out on both ends for them doing that. It gives us a tool for the study and gives the tool to be used by the patient, and it goes to what they were saying about having that tool in the field for patients to track themselves. The research stuff is usually far beyond what a

Page 243

patient really wants. They really want this information ahead of time, and they're prescreening themselves with it.

DR. GEWANDTER: I think we could mention that more simplistic measures are better in terms of their -- things that could be applied in the clinic as well as research so we can have some crosstalk is a good thing, and maybe incorporating it using technology so people can do it in their everyday lives as well.

We could consider mentioning that in the paper, that advocating for that in the future is a good thing. I don't see any reason why not to do that.

Does anyone have any other comments related to that or in general?

(No response.)

DR. GEWANDTER: Okay. Well, thank you-all so much for coming and for participating so well in the meeting. We want to again thank Valorie and Andrea for putting this together because, of course, we could not do any of it without their

Page 244

help. Thank you.

(Applause.)

Adjournment

DR. GEWANDTER: I also wanted to just thank Bob and Dennis for helping organize -- they're not listening at all.

(Laughter.)

DR. GEWANDTER: Thank you.

(Applause.)

(Whereupon, at 3:20 p.m., the meeting was adjourned.)

B

C

D

E

I

J

M

P

Q

R

T

U

V

W