UNITED STATES FOOD AND DRUG ADMINISTRATION

- - -

ADVANCING THE DEVELOPMENT OF PEDIATRIC THERAPEUTICS WORKSHOP

- - -

Thursday, April 16, 2015

FDA White Oak Campus, Building 31, the Great Room

White Oak Conference Center

Silver Spring, Maryland 20993

The meeting was convened at 8:00 a.m.

MEMBERS PRESENT:

DONNA GRIEBEL, M.D.
TERESA BURACCHIO, M.D.
SARRIT KOVACS, PH.D.
ELSA SHAPIRO, PH.D.
ANDREW MULBERG, M.D.
JONATHAN GOLDSMITH, M.D.
SUSAN WAISBREN, PH.D.
DAVID DAVIDSON, M.D.
ANN BARBIER, M.D., PH.D.
MELISSA HOGAN, J.D.
JONATHAN MINK, M.D., PH.D.
FLORIAN EICHLER, M.D.
ALISON SKRINAR, PH.D., MPH
GERRY COX, M.D., PH.D.
ELEKTRA PAPADOPOULOS, M.D., MPH
MELISSA PARISI, M.D., PH.D.
PETER COMO, PH.D.
KATHLEEN DELANEY
HEATHER ADAMS, PH.D.
JOANNE ODENKIRCHEN, MPH

P R O C E E D I N G S

[8:05 a.m.]

OPENING REMARKS

DR. GRIEBEL: Okay, everybody, I think it's time to get started. Good morning. I'm Donna Griebel from the Division of Gastroenterology and Inborn Errors Products and CDER at the FDA, and I'd like to welcome you all today on behalf of the FDA and the Steering Committee for this workshop, to the Workshop on the Assessment of Neurocognitive Outcomes in the Inborn Errors of Metabolism, which I'll refer to as IEM. Today's topic for the workshop is an important and challenging one, that is key to successful drug development for this rare disease field, which as I've said, I'll refer to as IEM.

Many IEM diseases have neurocognitive manifestations, and we learned last year at our FDA patient-focused drug development meeting on IEM, that these manifestations can range in severity from relatively subtle manifestations to profound deficits. And that even the more subtle deficits from an observer perspective, can have a marked impact on a patient's ability to optimally function in their day-to-day activities. Further adding to the complexity of the field, beyond the range of types of manifestations and their severity, the ages of the patients with these diseased ranges widely from infants to

adults. It's critical that we identify and define the deficits associated with the specific diseases, and that we develop methods and tools for assessing neurocognitive outcomes that can meet the challenges posed by this variability. In order to optimize our ability to bring forth to the market safe and effective drugs for these rare diseases, which have a profound impact on patients and their families. We anticipate that today's workshop will just be one -- the first one of a series of workshops dedicated to this topic. It became clear during this -- various Steering Committee meetings in preparation for this workshop that there's a lot to be talked about, and one day is just not enough.

We have an outstanding agenda today, and I want to take this opportunity to express my deepest gratitude to all the people who contributed their precious time and their skill to developing today's meeting. The Steering Committee consisted of representatives from academia, from industry, from NIH and multiple centers and divisions across the FDA: from FDA's Center for Drug Evaluation and Research, or CDER, there were staff from my division, as well as the Division of Neurology Products, the Division of Psychiatry Products, the Division of Pediatric and Maternal Health, as well as staff from the Study Endpoints and Labeling Development team. In addition, on the Steering

Committee there were representatives from CBER, the Center for Biologics Evaluation and Research, and CDRH, which is the Center for Devices and Radiologic Health. The industry rep was nominated by pharma, and I particularly want to recognize the hard work and expert contributions of five key people, who were key to the successful development of this meeting. First and foremost, Dr. Teresa Buracchio, who was the FDA lead for the entire effort from the FDA standpoint; Dr. Peter Como from CDRH; Commander Matthew Brancazio from our division; Dr. Elsa Shapiro, from University of Minnesota; and Dr. Susan Waisbren from Boston's Children. Before I bridge to Dr. Buracchio's talk, I'm told that I need to let you all know that it's of key importance for you to, sometime this morning before lunchtime, go out and order your box lunch, because the capability of obtaining food here is very limited, [laughs] so it's going to be important for you to go do that.

So now I'm going to turn the microphone over to Dr. Buracchio. She'll give a workshop overview and a high-level orientation to the regulatory framework within which we work in the United States to develop drugs for ultimate approval for marketing.

WORKSHOP OVERVIEW & U.S. REGULATORY REQUIREMENTS FOR DRUG APPROVALS

DR. BURACCHIO: As Donna mentioned, I'm going to provide a general overview of the regulatory considerations for drug development in the Inborn Errors of Metabolism, and then at the end I'll also provide an overview of the day ahead. I -- as far as disclosures, I'm an employee of the FDA and I have no conflicts of interest to report. And why I'm here: I'm a neurologist. I actually have fellowship training in Alzheimer's disease and dementia, but I have a clinical interest in cognitive disorders across the lifespan. I'm currently a medical reviewer in the Division of Neurology Products for Alzheimer's disease and epilepsy drugs, but I have previously served as a medical reviewer in the Division of Gastroenterology and Inborn Errors Products for the Inborn Errors of Metabolism.

So since we have a diverse audience today, we thought that it would be good to just lay some groundwork on what are the general principles for drug approvals in Inborn Errors products; although, this really applies to all products across diseases. So in order to approve a drug, the regulations require that the pivotal trials provide substantial evidence of effectiveness or clinical benefit, and there's two concepts

here. One is the substantial evidence piece, and the other piece is the clinical benefit, and I'll just touch on the clinical benefit briefly. So in order for a standard or full approval, a drug must demonstrate a clinically meaningful benefit. This could be an improvement in survival, or a benefit that is detectable by the patient, such as a clinical endpoint that we often define here at the FDA as a "direct measure of how a patient feels, functions, or survives." Or the endpoint could also show that you're decreasing the chances of developing an undesirable condition or disease.

Now, back to the substantial evidence piece. Substantial evidence of benefit requires adequate and well-controlled studies. The studies should be designed well enough to distinguish the effect of the drug from other influences, such as spontaneous change, placebo effect, or biased observation. The usual standard for approval is two adequate and well-controlled studies.

The definition for adequate and well-controlled studies is laid out in the regulations, and the major elements of these are: clear statement of the objectives of the study and clear statement of the methods of analysis, the identification of subjects to ensure that they have the disease or condition being studied, and assignment to treatment or control groups

that is comparable between the groups. The study design should allow a valid comparison with a control group to allow assessment of treatment effect. The control group -- we don't say specifically what the control group must be. It could be placebo, active control, dose comparison, or historical control. However, it is required that there must be a control group. Uncontrolled studies are not acceptable as the basis for approval. There should also be adequate procedures to minimize bias, such as blinding, and outcome measures must be well-defined and reliable, and an analysis of these outcome measures must allow for the assessment of drug effects.

So certain study designs can be vulnerable to bias. In particular, reliance on subjective measures in the context of an open-label study can lead to bias, such as expectation bias, in which a patient or the care provider know that the patient's on-drug and there is some expectation of benefit, therefore they may perform more well or put more effort into their testing than they would otherwise. There's also potential for bias in retrospective or chart-review -- retrospective chart reviews or historical controls. In this case, it's -- you know, usually unintentional, that -- but it's selective entry of data -- selective collection and entry of data into a chart, and when a provider decides what to record and what not to record, and then

later someone comes and tries to extract that data and fit it into clinical measurement scales.

So, due to these tendencies to bias -- being prone to bias -- historical controls should usually be reserved for special circumstances. For example, in studies with high and predictable mortality, or studies in which the effect of the drug is self-evident.

There is another pathway for approval called the Accelerated Approval Pathway. This falls under Subpart H for drugs and Subpart E for biologics. In this Accelerated Approval Pathway, it's generally intended for diseases who may have a long latency period, or for which there is an extended period of time before a clinical benefit can be demonstrated. These regulations allow drugs for serious conditions that fill an unmet need, based upon a surrogate endpoint.

A surrogate endpoint used for Accelerated Approval is a marker, such as a lab measure, a radiographic image, a physical sign or other measure, that is thought to predict clinical benefit, but is not itself a direct measure of clinical benefit. Likewise, and intermediate clinical endpoint is a measure of a therapeutic effect that is considered reasonably likely to predict the clinical benefit of a drug, such as an effect on irreversible morbidity or mortality. So if a drug

does get approved under Sub-part H or E, with a surrogate marker, the clinical benefit requirement still is in play. It may not be required at the time of approval, however it is required that clinical benefit must be verified and described in a study. This may be a continuation of an ongoing study, or it may be a separate study, but the study should generally be under way at the time of approval. And it should be noted that this can be quite challenging, because once a drug is approved, it's difficult to get patients to enroll in controlled studies.

And of course, we do assess safety throughout the course of development. There should be adequate safety to perform a risk-benefit analysis prior to approval, and safety is continued to be monitored after marketing. FDAA does allow the FDA to require post-marketing studies, to further assess safety.

Clinical trial design can be challenging in rare diseases and specifically in the Inborn Errors of Metabolism. The most obvious reason is of course, that these are rare diseases, so there are few patients available for studies. You don't usually get the chance to do multiple trials, so it makes getting development right critical from the start. These are usually serious, life-threatening diseases with high unmet medical need, and one of the big challenges is the heterogeneous clinical presentations of many of these diseases. There can be

a great deal of variability, both across patients and within individual patients, who may show variability of the severity of their symptoms over time. The natural history is often not well-characterized in these diseases, and well-defined endpoints or biomarkers can be lacking.

There's also challenges to interpreting treatment effects in these trials, so reliance on data from open-label trials or historical control trials, requires that the treatment effect be large enough to interpret in the absence of a placebo-control, or to overcome inherent bias that might be in the study design. The interpretation of outcomes will be challenging if the treatment effect is not dramatic, the effect is a slowing of disease progression with small differences of unclear clinical benefit, or if there is substantial interpatient variability in treatment response.

To address some of these challenges, the program should be based on solid scientific foundation. It is very important to have an understanding of the drug mechanism of action and the disease pathophysiology. The natural history of the disease must be characterized. The study design must be based on the population under study and the expected drug effects. And substantial evidence of effectiveness can be shown by a single study, but results shown must be statistically

persuasive. And there is a guidance for effectiveness that lays this out a little more clearly. Generally the idea is that there should be a large treatment effect. There should be a strong, statistically persuasive result, such as a very small p value, and results should be consistent across multiple supportive endpoints.

Before initiating a pivotal clinical efficacy trial, it's critical to map out the clinical development program, conduct a natural history study early in development, and to design efficient early-phase trials that can inform the design of the pivotal efficacy trials. And of course, we do assess safety and tolerability throughout the entire drug development process.

So I think the key message you're going to hear today is that you need to think about your development plan very early in [laughs]. As you're starting out, begin with the end in mind. So today's meeting, we will be covering: Morning session, we'll discuss natural history studies. We will be discussing in Session 2, what is efficacy for neurocognitive outcomes? We will also be touching on presymptomatic disease with this. The afternoon session, we will talk about choosing, selecting scales and developing disease-specific scales. And the late afternoon session, we'll talk about methods for standardizing cognitive

assessments. And as Donna said, the assessment of cognitive outcomes in Inborn Errors is a very large topic, and obviously we can't cover it all in one day. The goal for today is to start a conversation on how to approach cognitive assessments in this population. And although the focus is generally on neurocognitive endpoints, we do understand that these diseases usually affect multiple organ systems, and so we may not limit ourselves solely to cognition in our discussions today. We may talk about diseases more broadly. And, very important to point out, that we know that there are many, many Inborn Errors diseases out there, and we only have a few to talk about today. Not to slight any other diseases, but we just found a few good examples that we thought would be representative and might be able to be applied across a variety of disease. So we hope that these conversations can be used broadly, and to guide other -- you know, to guide other Inborn Errors diseases or other genetic diseases.

So thank you and welcome. I think we have a really great workshop lined up for today, and hopefully, we'll have some really interesting discussions.

So next, I believe, is Sarrit Kovacs, who is from the Study Endpoints -- I'm not sure how to get to the next one

[laughs] that -- are you doing it over there? Okay. So Sarrit Kovacs from the Study Endpoints and Labeling Division.

ROADMAP FOR ENDPOINT DEVELOPMENT

DR. KOVACS: Good morning. My name is Sarrit Kovacs, and today, I'll be presenting you with a roadmap to endpoint development. I'll be presenting my own views, which do not necessarily represent an official FDA position. I work as a study endpoints reviewer in the Center for Drug Evaluation and Research at the FDA, and I'll be providing the FDA perspective on the roadmap towards endpoint development, specifically clinical outcome assessments in clinical trials.

I'll be covering a few topics related to instrument development. I'll give you an overview of the development of clinical outcome assessments, including PRO measures, present a few ways in which the agency can work on instrument development with stakeholders, and give you an overview of FDA's Drug Development Tool, or DDT, qualification program: a way to work with the agency outside of an individual drug development program.

The FDA's code of federal regulations states that it's necessary that drug developers document substantial evidence of treatment benefit from adequate and well-controlled clinical trials, as Teresa explained earlier. The regulation also specifically indicates that the methods of assessment of a

subject's response should be well-defined and reliable. This is important to the agency when reviewing a PRO measure. In other words, well-defined and reliable are the key criteria by which the FDA evaluates clinical outcome assessments to document evidence of treatment benefit.

The purpose of PRO measures or clinical outcome assessments are to aid in determining whether or not a drug has been demonstrated to provide treatment benefit to patients. Based on the regulatory standard of well-defined and reliable, the FDA finalized a PRO guidance for industry in 2009, providing industry and instrument developers with detailed guidance regarding how to establish the content validity of a measure. Content validity refers to how well a measure -- or an instrument -- measures what it purports to measure. In establishing a measure's content validity, it is essential to include input from the target patient population. Patient input can be obtained via qualitative interviews and focus groups, to identify the core signs and symptoms important to patients within a specific disease or condition. It's also important to confirm that the items in the developed measure are relevant, clear, and understood by those patients. The PRO guidance also includes guidance regarding the quantitative evaluation of a measure: its psychometric properties, such as its reliability,

construct validity, and ability to detect interpretable and clinically meaningful change. All of this work is to ensure that the newly developed PRO measure in fact meets the regulatory standard of well-defined and reliable. All clinical outcome assessments need to adhere to good measurement standards, in developing them and documenting that they are well-defined and reliable requires a lot of work. However, it is important to be able to determine, in a failed clinical trial, whether it was in fact the drug that failed, or if it was simply an unreliable or invalid outcome assessment that was incapable of detecting a clinically meaningful treatment benefit.

The PRO guidance provides an optimal approach to PRO development. However, flexibility and judgment are both necessary in order to meet the practical demands of drug development, such as tight development timelines. In addition, the FDA encourages drug sponsors to engage in early and continued communication with the agency during instrument development and evaluation.

The Study Endpoints team developed a roadmap to patient-focused outcome measurement and clinical trials, to help instrument developers develop or prepare for the development of an outcome measure. This version of the roadmap is a pared-down

version, which does not include all of the sub-bullets under each letter category found on the FDA website. I'm not going to walk through this whole slide as it is on the website, but I'll make a few important points. This roadmap describes an orderly process. It is especially helpful to instrument developers when they plan to develop a measure in diseases or conditions that have not yet been adequately conceptualized in terms of how patients feel or function.

The first step shows that a developer must first understand the disease or condition of interest, before beginning to develop the measure. The second step shows that one must identify what aspect of treatment benefit is of interest to measure, such as how a patient may feel or function. After the first two steps are completed, the developer can move on to step number three, and select or develop the outcome measure.

There are three main types of outcome assessments that can be used to evaluate patients' treatment benefit. A patient's survival, which is appropriate in contexts where the disease or condition impacts survival, such as cancer. In addition, there are surrogates which are objective and to not rely on human judgment, such as biomarkers like blood pressure. The third type of outcome assessments, which is most relevant to

this presentation, is clinical outcome assessments. There are four types: outcome measures where the reporter is a patient, clinician, or other observer; and performance outcome measures. PRO measures are important in evaluating patients' treatment benefit, and are extremely useful to directly assess how a drug may impact how patients feel and function in their daily lives. PRO assessments can be used when patients can reliably and validly report on themselves. However, in some patient populations, it may make sense to use another type of clinical outcome assessment. For example, a very young child or someone with cognitive impairment may not be able to report for themselves, and likely an observer-reported outcome measure reported by a caregiver could be used to evaluate treatment benefit.

A clinician-reported outcome measure is most appropriate when clinical judgment is required to interpret an observation about a patient. Finally, when a patient's ability to perform specific tasks is useful in determining treatment benefit, such as walking a certain distance within a specified time limit, a performance outcome assessment could be the appropriate clinical outcome assessment to develop and use.

There are two ways in which the FDA can work with stakeholders to either evaluate existing tools, or develop a

novel clinical outcome assessment. The first way is the traditional way, or within an individual drug development program. This is where the drug sponsor submits an IND, or Investigational New Drug, application to the agency before conducting any clinical trials. The agency encourages drug sponsors to meet with the medical review divisions early, even in the pre-IND stage, to discuss instrument development. After a drug sponsor establishes the content validity of a measure, the sponsor can include the measure as an endpoint in a Phase-II clinical trial to test its psychometric performance. This way, if the psychometric properties of the measure look acceptable in Phase II, the sponsor can feel fairly confident that the measure will be an adequate measure of treatment benefit in pivotal Phase III trials. If the measure is included as a primary or key secondary endpoint in Phase III trials, and it adequately detects a clinically meaningful treatment benefit, it may be suitable for inclusion in labeling claims.

The second way in which the FDA can work with instrument developers, is a newer process falling outside an individual drug development program. This second way is within CDER's Drug Development Tool Qualification Program, and is intended to produce qualified measures for use across multiple development programs. Within this program, the agency works with

many stakeholders, including consortia, patient groups, individual academic investigators, and drug developers to develop and qualify outcome assessment tools, making them publically available across multiple development programs within the appropriate context of use.

To help explain the qualification program process, the Study Endpoints Team created the wheel and spokes diagram that is displayed on this slide, which is also a pared-down version that doesn't include all the sub-bullets that is found on the FDA website. Before the agency reviews a clinical outcome assessment for qualification, the submitter will need to first identify the targeted context of use or patient inclusion/exclusion criteria, clinical trial design, and hierarchy of endpoints in the trial. The submitter will also need to identify the measurable concept of interest, or measureable way to detect treatment benefit. For example, how a patient functions when engaged in activities of daily living.

Second, the submitter will need to establish the measure's content validity using patient input, as I mentioned earlier. Once the measure's content validity has been documented, the submitter can move on to Spoke Three, and cross-sectionally evaluate the instrument's other measurement properties, including evaluating the scoring algorithm. At this

point the submitter can consider submitting the measure, its user manual, and the qualitative and quantitative evidence to the FDA for qualification for use in exploratory studies. If the submitter would like to have the measure qualified as either a primary or a secondary efficacy endpoint to support labeling claims, longitudinal evaluation of the instrument's measurement properties and interpretation of clinical meaningfulness must be examined and documented. Even after a measure is qualified as an efficacy endpoint to support labeling claims, the measure can be modified for a new context of use. During this process, the item wording or response options may change, and the measure may be translated or culturally adapted for new populations.

To further aid the submitters or instrument developers in navigating the Drug Development Tool Qualification Program, the FDA published a guidance for industry and FDA staff in 2014. In this guidance, the agency describes the process for qualification of DDT's, such as PRO assessments or clinical outcome assessments. The qualification program is a process used by the agency to work collaboratively with interested parties or submitters, to develop qualified measures of treatment benefit for use in specific diseases outside of an individual drug development program. Qualified drug development tools or measures are deemed by the FDA to have a specific

interpretation and application in drug development and regulatory review, within the qualified context of use or target patient population.

It’s important to note that clinical outcome assessments used in clinical trials are not required to be qualified through this program; however, one assessment are developed collaboratively, in consultation with FDA, and then ultimately qualified drug companies can feel confident that the agency agrees with the content and measurement properties of an instrument, which aids these companies in pursuing drug development in important therapeutic areas.

Here are some links to the PRO guidance, to the qualification program webpage, and the qualification program guidance. Thank you.

[applause]

SESSION I:

NATURAL HISTORY STUDIES IN IEM

DR. MULBERG: Good morning. I'm Andrew Mulberg, the division deputy in DGIP and happy to chair this and the next two sessions with Elsa Shapiro, who will introduce herself. And I just want to welcome everyone and once Elsa introduces us, we'll introduce our first speaker for this next session on Natural History Studies and Inborn Error of Metabolism.

DR. SHAPIRO: [inaudible]

DR. MULBERG: Press the red button.

DR. SHAPIRO: I'm Elsa Shapiro from the University of Minnesota, and I'm very pleased to be here today as a non-FDA person.

[laughter]

DR. MULBERG: I take that as a compliment, I suppose [laughs]. But we're all here to really partake in the briefing.

DR. SHAPIRO: For the same purpose, right.

DR. MULBERG: Right, exactly. So Dr. Jonathan Goldsmith is part of our rare disease program, the Office of New Drugs at FDA. He will discuss developing naturalistic studies for rare diseases. And Dr. Goldsmith, welcome.

DEVELOPING NATURAL HISTORY STUDIES FOR RARE DISEASES

DR. GOLDSMITH: So good morning. I want to talk to you about drug development but in the context of interdevelopment of natural history studies for rare diseases. Inborn Errors of Metabolism are rare diseases, so they fit within this context, the way that I look at things [unintelligible]. I wonder if these buttons will work. I think I've got it.

MALE SPEAKER: I think it's this one, this button.

DR. GOLDSMITH: That one. I don't have any conflicts of interest to report. I'm an employee of the Food and Drug Administration. I'm the acting associate director of the Rare Diseases Program in the Office of New Drugs in the Center for Drug Evaluation and Research at FDA. I'll tell you about my background. I'm a hematologist by training. I've had a career-long interest in rare diseases, as that's what hematologists do. And I have experience in drug development in an academic side as well as in regulated industry and in government and at the NIH. So I had a chance to look at this from several different aspects, and it's given me some perspective, but I think I'm pretty useful in the long run.

The opinions that I express today are mine and not those of the Food and Drug Administration. So let me tell you a little something about the Rare Diseases Program here at the Center for New Drugs. Our mission statement is “to facilitate, support, and accelerate the development of drug and biologic products for the treatment of patients with rare disorders,” which gives us a kind of sweet spot for me in terms of what I like to do in my daily work. You try and help people from all aspects of drug development to try and move their products forward to maybe improve the public health.

We have some objectives that we follow to try and accomplish this particular mission. We want to coordinate to development of [unintelligible] policy and procedures and training for the development of treatments of rare diseases. We want to assist outside development and maintenance of good science and this involves interactions with clinical investigators as well as regulated industry. We want to work collaboratively with external and internal rare disease stakeholders and to maintain collaborative relationships with [unintelligible] review divisions to promote consistency and innovation in review. To maintain a thoughtful approach that may help bring products to market.

There's often confusion, I think, between what's a natural history study and what's a registry. I want to spend a little bit of time trying to clarify that because at the end of the day, I think what your goals will be to develop natural history studies, and not registry, although they can play an infrastructural kind of beneficial role. AHRQ and others -- you can find these kind of definitions around -- that a registry is an organized system, uses observational methods to collect uniform data to evaluate specified outcomes for population defined by a particular disease, which sound pretty good on the surface; it sounds like that's what you're thinking about, but actually there are some deficiencies and I just want to highlight some of these to help you think about how it's different from a natural history study.

The common data elements that are used in a registry really need to be defined and validated, and they very often are not, to be honest. The data sources that are used are often patient records. It is not an instrument that's been designed for the purpose of data collection of focused on the answers that you're seeking. And as a result, it doesn't give you the kind of information that's supportive at the end of the day for regulatory work that you're going to have to do down the road to get your drug to market.

There may be communications that come from sources with varying levels of interest and expertise. Does each practitioner try to fill out some kind of a form for you? It's not going to be very useful for a registry. And yet that happens all the time. And I think we all know that. The people that extract the data have varying levels of expertise. They may be a good output or they may not. They may be a full-time employee, they may be a volunteer. It may be done after hours, so there's a tremendous variation in registry in terms of qualities; however, registers do play an important role for regulatory point-of-view because they can help fulfill post-marketing requirements and commitments. Registries can be put in place to do safety file work. And there an instrument that we accept all the time in follow-up. And they can also do a longer term assessment of an intervention, in which a firm may have limited resources and this is the best they can do sometimes to show that a drug brought to market is actually doing what it's supposed to be doing. They also can serve as a lead into natural history. I think they can serve an important infrastructural role. Registries are often done in academic settings, and some of the information can be used and it can be extracted to be used, as you want to build a natural history study.

So now, to move to the definition of a natural history study: This is borrowed from NIH, from their rare diseases program, and modified a little bit. So the definition would be that the natural course of a disease, from the time immediately prior to its inception, progressing through its pre-symptomatic phase and different clinical stages to the point where the patient is either cured, chronically affected or dead without external intervention.

And I think it's a definition that most people, I think, accept at this point in time. One of the key things about the characteristics of natural history studies is that they are a subset of registries, is the way I think about it in this sort of stick diagram here helps you understand that it's a Venn diagram. It's not -- natural history studies are part of a bigger universe, but there a much better defined part of that universe. They have better definitions and they have better utility at the end of the day.

When you design a natural history study, it should have a specific purpose -- to improve the understanding of specific disease. Natural history studies are prospectively planned. Their designed to be comprehensive and detailed. And this is really the critical part because they describe the disease independently, as specific observers of the clinical

sites and interventions. People who do review of drug development often see what's called center effects. The drug works in center A, but doesn't work in center B. And there are a lot of explanation to that. I think that's probably a whole days discussion in itself but a natural history study has to get outside of that. It has to be able to be something that can be applied broadly, and it will then support your contention about the development of your new drug and the impact on the health of the affected population.

Natural history studies should involve all stakeholders. They should be widely available so if its developed in academics of its developed by regulated industry, it should support the development of drugs for that particular disease process and it should be something that's almost in the public domain because it will help -- you know, it's the rising tide that lifts all boats, and I think there's a real urgency for people to think about this as they develop natural history studies, that there's a greater good beyond their own particular drug development.

These natural history studies are intended to describe the disease under study and if done properly, they can provide important information that will help shape drug development and help identify clinically meaningful end points, which is really

to goal of what you're trying to do when you develop your natural history study.

There are a few different types of natural history studies, and I'll just briefly go through these so-called retrospective studies. Unfortunately, these may be of lesser utility because they may often be incomplete and hard to interpret because the terms weren't agreed upon prospectively. And when one investigator say X, they really mean Y, and so on. Historical controls, which are part of the development of retrospective natural history studies comes out of that. They may lack important prognostic co-variance. They have unknown or unrecorded historical data, and if you do a retrospective natural history study, people think you just sit down and do it. No. It needs to be prospectively planned, even if it's retrospective, to get the maximum value out of this, you have to do planning. It isn't something that just falls off the pages of a chart.

This [unintelligible] is to do a prospective trial, and a prospective trial is a well-planned trial -- it's a trial that looks at the various aspects of what your planning and may even include a pilot, in which you have kind of an alpha version of even a beta version before you actually go out and do your

natural history study and collect the date that you will use in the long-run to support your drug development program.

There is also cross-sectional natural history studies. These are of interest when you have populations with large numbers of subject, large numbers of participants and trials, potentially. And it looks at the population and there's an assumption that people will be at different stages of the disease or have different clinical manifestations of the disease. We just cut a wide swathe across the patient population so we get to see everything, and it's interesting. It's useful, but it's static. It's not a dynamic assessment, and in the long run I think you'll be better off if you actually invest your time and effort in developing prospective longitudinal study in which you can actually track individual participants from time A to time B to time C and it will have the most value for you at the end of the day as you try to develop your natural history studies.

If you select A, elements for natural history studies, you should think about within the context of the disease. If you're actually going to be doing this study for -- and think about the following feature. As we've already heard, some of the most important clinical relevant finding, or the substantial

day-to-day impact, that are of greatest importance to patients and families.

And how do people feel, function, and survive? And these could be simple things. I talked to some families from a spinal muscular atrophy group, and simple things can be really clinically meaningful. We don't always think about this. We're always trying to think about genetic causes and biochemistry, and they talked about the fact that their young child, between 2 and 5, was able to sit up, handle secretions, and have a face-to-face conversation with the family. That is a clinically meaningful benefit. And so try to think along those lines as well as along some of the more traditional lines that we think about. We also have to have data elements that explore the potential moieties of the studies -- of the disease state and the prognostic characteristics and the disease features have to be known to help formulate a sensitive clinical endpoint. And the natural history study will actually, if you design it correctly, will identify some of these features and you'll find out what are the critical end point and you'll find out something about the timing of the development of the clinical endpoints, which will help you in prognostic development.

What makes drug development for rare diseases more difficult? Okay, well as I think you all know rare diseases are

in general poorly or incompletely understood. Even the world authorities who have seen, you know, 23 cases of a disorder that affects 500 people, there pretty well schooled on it, but there are lots of defects, and if you actually drill down into what's published, there are lots of holes in terms of what goes on over time, and that we don't really have enough evidence, enough data, to make decisions that we need to make as we design studies and try to choose endpoints.

If a disease is extremely rare, and there is very little understanding. If you have a small disease-state population, it also may restrict study design in the ability to replicate trials. As you just hear earlier, the pathway to licensure is to automatically put in controlled trials, to phase III studies. Well, you have a small population you are lucky to do one study like that. So you have to really understand the population, the natural history as well as possible because you're going to have to develop a compelling body of evidence that will convince the medical reviewers that this drug actually works.

There are also problems that complicate this. There's phenotypic diversity within specific disorder and genetic subsets are now becoming more and more prominent. And what is there impact on the natural history studies? Also, when you're

dealing with rare diseases, there's no rut in the road. Nobody has already developed five of six agents like asthma that has 800 agents for the treatment of that disorder. And if you try and bring an asthma drug to market, there's a pretty clear way to do it. And you can pull out all the packaging from all the drugs that have been licensed in the last 20 years and you can look and see what they did and what the claims are and you can bottle yourself like that and you can meet with the review division and they'll tell you "this is what you have to do." Here, it's not like that. So the more evidence you have from a natural history study and the more you know about the population. The better equipped you will be to actually get that drug moved through the process to market.

Okay, so natural history studies for rare diseases these help inform drug development, they facilitate the design and content of adequate and well-controlled trails, and they also give some scientific basis to this to have a good understanding of the pathology of the disease and the natural history of the disease as we're talking about that as a matter of credibility to the endpoints that you select. It also provides information needed for drug development for poorly understood, rare diseases and can support the development of a bio suppository, which may be useful.

What are some of the challenges? It's difficult to define the disease population sometimes. You have to understand the full range of the disease manifestations and the subtypes if you really are going to develop -- I think that will come out of your natural history study but then it may help you refine it later on. But in the front end, you have to understand that there's [unintelligible] diversity. There's also diversity in terms of day-to-day changes and severity for a particular patient and you have to take that into account. You need enough data point that you can figure that out in terms of the longitudinal follow-up.

There's also interpatient variability. You know, you think about monogenic disorder like sickle-cell disease; there's tremendous phenotypic diversity. It isn't just straight in a line. And you have to be able to figure that out and collect the data that you need to actually understand this. The steady duration is the big impediment. You've already seen the slide about the end; think about the end at the beginning. It takes a long time to develop such a study. I wouldn't initiate a natural history study when I was beginning my phase III trial. That would be way too late in the game. And when you should think about this is when you're doing preclinical work and have that good idea about a therapy. That's probably the time to

begin to think about doing this work and how would we actually develop a natural history study. And if you move from an academic setting with a great idea for a therapeutic and you enlist a corporate partner, as soon as you enlist the corporate partner, that's probably one of your first items of business because we need to understand this. I think most of the people who are serious in the rare-disease space get it. And if they're willing to invest in this kind of an exercise, if you will, it will help bring your drug from the beginning through the process and to market.

So keep in mind this is a long process. It can take several years. Some of these registries have evolved into natural history studies are like ten year efforts. It's probably too long for practicality but keep in mind that that has happened. One of the problems because of this long time duration is that the standard of care can change while you're doing your natural history study. Something else develops, they developed a better simple thing like more daily fluid something else, there's some other intervention that comes along. Well, if that happens, you need to turn those lemons into lemonade. And you can do it if it's a well-designed natural history study you can see when the intervention occurred, when it entered clinical practice, when people entered onto the new care. You

can modify your data collection and then take that into account as you move forward. And it isn't the end of your natural history study, it an addendum to your natural history study.

Also, the choice of subpopulations maybe very important because there are -- people talk about enrichment to trial populations and that's possible -- it's harder doing rare disease, fewer subjects. But to think about the people who might potentially be more responsive to a proposed therapeutic and could you select some of them from what you've learned in the natural history or some of your genotyping studies, let's say. And that may be an important bit of information to add.

[unintelligible] some opportunities, then, if you perform a well-designed natural history study and you conduct it long enough, you'll be able to develop and select outcome measures that are more specific or sensitive to changes in the manifestation of the disease. And on the kind of more negative side, [unintelligible] you can more quickly demonstrate that there are safety and efficacy concerns, which means you could end the developmental program before it goes too far. So it gets to review a positive and a negative effect and from a corporate point-of-view, it's useful to know that something has a safety signal early in the game because it may actually change the developmental process.

There are less opportunities to perform a risk/benefit analysis. This is something that is standard FDA fare. Everybody, every reviewer looks at risk/benefit analysis and that's something that we will be informed effectively by performing a natural history study. You can also help identify and evaluate biomarkers. You can develop new or optimized biomarkers that may provide a proof of concept and guide that reception for the recognition of safety issues. And also the other product biomarkers is while your developing these, you may also be able to validate the predictive value of a proposed biomarker, including the measurement technology themselves. It's not an issue that becomes sticky towards the end of development. How good is the test? How do you know the test works? What's the anal like? What are the controls? And that's something that can be done during the course of some of these natural history studies. It's a little bit on the side, developing biomarkers, but it's a popular area and if these are validated by a markers, and they could be validated by the natural history study, then they've proved their useful in terms of drug development.

I am going to give you an example of something that was successful recently. There's a recent natural history study that was used to support drug development last year that was a

drug U.S. [unintelligible] alpha that was licensed for Muco-A syndrome, mucopolysaccharide IVA. And what the firm did is they got access to data that had been collected over a long period of time. Looking at six-minute walk test information on individuals who were affected by this disorder and they compared it to healthy reference values. Six-minute walk test is a test of exercise capacity, which some people call endurance. And basically, if you look at the chart, over time, by ages 12 to 18, there were 75 percent less endurance in affected individuals with this particular disorder. When they then did their study with their intervention, they were able to show that endurance was preserved, and that it was radically different from what this natural history study had demonstrated before. And it supported the licensure of the product, basically. It was licensed about a year ago, I think.

Okay, so I'll try to give you a quick overview of natural history studies. They're very complicated. But their also incredibly useful. They are hard to set-up, but I think the rewards fat exceed the investment of time and effort. And so I would encourage everybody to keep this on their radar screen and I'd be glad to answer questions later on or during the panel. So thanks very much.

[applause]

DR. MULBERG: Thank you, Dr. Goldsmith. You're welcome to join us here at the panel now or later. Next speaker will be Dr. Elsa Shapiro, who will be speaking on challenges of cognitive and behavioral testing in in born errors of metabolism specific natural history studies.

CHALLENGES OF COGNITIVE TESTING IN IEM NATURAL HISTORY STUDIES

DR. SHAPIRO: Well, I'm very pleased to be here today and to share with you some of the experiences I've had in the natural history studies that I've been involved in. Here's a brief biography, but actually I'm going to talk a little bit about what I've done in the field. So I'm going to begin by talking about the purpose of natural history studies. I'm going to then talk about what we're measuring and what the challenges and guidelines are for neurocognitive testing. There's a lot of challenges, a lot of solutions. I'm not going to be able to cover it all. These are very dense slides, so looking at the slides, probably they will give you a lot more information than what I'm going to say, so there's a lot here to ponder. As we know, Inborn Errors of Metabolism are rare disorders. Most effect the development of the central nervous system with a childhood onset with progressive degeneration and many of them have what I define as a childhood dementia, which I'm going to talk a little bit about that a little bit later in my talk, what that is and how it differs from an adult dementia. So, I'm going to start by talking about my own personal experience. Starting in the 1980s, I've been doing this now for 30 years, and what some of the lessons are that I learned in doing this.

So, my -- the first lesson that I learned when I started doing this in 1982, is that one size doesn't fit all. It was the habit at that time to sort of lump all the mucopolysaccharide doses together, all the leukodystrophies together and, in fact, as I began to see these children, I began to realize that from a cognitive standpoint and a behavioral standpoint these were all very, very different from each other and each of them have some very specific things that affect their cognitive development that are unique and important and that they each needed a disease-specific approach.

The second lesson that I learned is that we needed to know how untreated patients differed from those that had been treated and that we needed to do natural history studies. Initially, I began with clinical evaluations and I tried to maintain a protocol at that point and to develop a protocol for each of these diseases but it was in a clinical setting and that was how we did it at that point. And so, I began to collect data personally about these diseases. And then I began to realize that the standardized scores and the test scores that were -- I was obtaining just didn't fit the patients very well and that I needed to find some alternative approaches because the tests, as they were developed at that point, did not meet the needs of the children -- the assessment of those children.

And I was working a lot at that point with children with Hurler syndrome. I'm going to use that as an example because that was the first treatment we had that was effective in making a difference in the lives of these children and that was hematopoietic cell transplantation. At that time it was all bone marrow transplantation. And, you know, I think that we need scores that are sensitive to cognitive change because there's now an explosion of new treatments that are going to demand precise natural history data; and at the beginning, it was not so precise.

So, what is the importance of these studies? One, is that the -- this has been said before -- the natural history is often unknown, we don't know the rate of disease progression, the age of onset, the nature of the disease process, and you need to know the rate of cognitive decline -- the rate, keep that in mind, of cognitive decline, to gauge the time clinical trial needs to run before you are going to see effects. And that's very important. The next thing that is important is that rare disorders have a very limited pool of subjects, so that every single patient, every patient is valuable in understanding the disease. And once the natural history is documented, it diminishes the need for multiple comparison groups. And as we

all know, sometimes we need to go internationally in order to get enough patients to have a reliable study.

So what are we measuring in these children? So let's talk a little bit about dementia in childhood. It's very difficult to diagnose because the onset can be subtle and it mimics developmental disorders, there's no pre-morbid history for comparison as there is in adults, and the rate of development is affected, as the brain is developing all through childhood. So you have a rate issue and I'm going to show you a model that I developed. Children are more affected by environmental and medical factors; they're very sensitive to these things and so you have to take these things into consideration as well. So this is a model that I developed here -- I wish I -- is there a -- yes, there it is -- so, when I was first working in this field, I developed this model, thinking this through theoretically, that you have a vector of normal development, you have a disease effect, and then you have a resultant and that resultant can -- many times looks like this. Where you have a child that is developing normally until the disease has an effect, where the child plateaus or slows in their development then declines. So this is their developmental status. This is their chronological age. And you get a curve that looks like this. And I began to see that the only way we

could get this kind of information was to use age equivalent scores to allow detection of slowing, plateauing, and declining abilities, which can't be done with -- I mean -- it can't be done readily with IQ measures.

So, let me give you an example with Hurler syndrome and what we have here is a rare disease with -- that effects every organ of the body with an average age of death without treatment of 5. And in 1983, the first transplant was done for Hurler syndrome in the United States in the -- in the institution in which I worked. This is the spectrum of disorder. You have a mild form progressing to a -- or not progressing, but a spectrum that includes an intermediate form and a severe form. With cognitive abnormalities found in all three phenotypes. And, this is just a diagram to show you all of the many organ systems that are involved in this disorder.

So this was the first graph that I constructed; this dates from the 1990s. And I was using the Bayley Scale of Mental Development to test these children. They were all between 0 and 4 years of age. And so this is what I was able to collect. These were either pre-treated or untreated children. At the time that this testing was it was untreated. And then I noticed this. Take a look at this. The lowest score that you can obtain on a Bayley is 50 and so you have a truncated

distribution, which doesn't really -- should look like this. A much steeper kind of decline and the -- I began to see that we couldn't use a mental development index that continues today because all of the test that we have in our armamentarium of cognitive testing have a floor of 40, 50, something like that, depending on the test. And, so, we're not getting the whole spectrum, we can't really see what happens when a child gets below that 50 mark.

So instead, I began to use age equivalence and this is very interesting. So this is old data and this was age equivalence on either the Bayley or Mullen collected before 1997. And you see a similar pattern to the theoretical pattern that I showed before. At the very beginning they approximate normal development and then there's a lot of variability, but then they decline at the end. Take -- just take a note of this line here, 28 months. This is the age equivalent -- no this is actually -- yes this is the age equivalent here. This is new data. This is from a natural history study that was just completed on Sanfilippo syndrome type IIIA, and this is a very similar kind of pattern that we're seeing. This is a longitudinal perspective natural history study in which we were able to document a similar kind of pattern. Take a look at this line right here. Not -- actually that line is in the wrong

place, I don't know what it should be -- it should be down here. It got moved in the process of translation probably from a Mac to a PC, I'm not sure, but that line should be down here. Okay. Just, in your mind's eye move it down a little bit. And so, we had only one subject in our rapid-progressing group that went below -- that was above 28 months. So, when you're doing a natural history study, you not only get an idea about the progression, but you get some scientific ideas too. What is it about these diseases that does not allow development beyond 28 months, except in rare cases? And so you have a lot of variability, but you also see that there is a sort of sealing of development in some of these diseases. It -- you know -- we've been mulling about that and we don't really understand it, but it's a serendipitous finding that may help clarify neuropathology at some point.

When we went ahead and did our transplants, we used a similar model and we used age equivalent scores and you can see we learned a lot from this. We learned that, you know, under 24 months that -- or, actually, this is under 18 months that the developmental curves -- these are individual patients -- improve over time. And for those transplanted later, they don't improve as much. They are stabilized, but they don't improve. So one of the things we found was that the pre-transplant developmental

level and their age at treatment was able to predict -- predicted their outcome and we've set a year of -- an age of 2-year as sort of the minimum -- the maximum that we're going to be doing treatment in these kids with transplant. Note that, even with transplant, these developmental trajectories are no normal and so this is an imperfect treatment and there's a need for better treatment for these children, even today.

So I want to move ahead to some guidelines for neurocognitive testing. First of all, it's very important to use a developmental model to provide growth information with an appropriate metric and sometimes mental development indexes don't work, or early learning composites, or whatever you have in -- as an outcome measure on standardized tests that are available in the field. This kind of developmental model that I'm talking about is most appropriate for young and for very impaired patients. I'm not talking now about 8-year-old children with Hurler-Scheie syndrome or a child with a slower progressing disease. This is the kind of approach that we've taken for very young and impaired patients. And we can construct development growth curves from raw scores or proxies of raw scores, such as age equivalence scores, which have normative data from the test development companies. Psychology's only recently understood this.

This is something that is missing in the field of psychology because pediatricians have known about using this approach for height, weight, head circumference for a long time, but we have not really understood this as an appropriate way to track children who are very low functioning or very young. And you can gain knowledge by this approach by -- about whether the child is still developing, whether they have plateaued, or whether they are in a downhill course. It also provides better statistical power with fewer patients, when you have multiple points for each individual patient. And that's very important in rare diseases. So you have to make a decision in a natural history study about whether you're going to use norm-reference scores, whether you're going to use standard scores, age equivalence for older, mildly to moderately impaired children. You know, norm reference scores are very important and they can be standardized scores, but as I said before, with the young children and children with severe impairment, the floor makes them insensitive.

So an alternative is to use age equivalence scores, which you can obtain on a test such as the Bayley or the Vineland Adaptive Behavior Scales, which I will mention a little bit later. And we found -- I will tell you now that we found a correlation of 0.94 between an age equivalence score on a direct

measure, such as the Bayley, and a parent observation, which is the Vineland. So that's pretty important. And I'll come back to that.

The second guideline is whether we should be -- or the second question is whether we should be using disease-specific or generic approaches and I think the important thing is to gather preliminary information about the phenotype and consider that phenotypes may vary with age of onset. You -- generally, a natural history study should be designed with the phenotype in mind and usually generic approaches are not useful in these rare diseases. And you need to plan how a natural history study may increase the specific knowledge about the disease. I really get upset when people use general terms like "mental retardation" or "behavior problem." Well, how is that child retarded? What does that mean? Have they at one time had a normal IQ and have deteriorated? Are they changing? Is that behavior problem -- what is the nature of that behavior problem? So it's really important to be specific.

So when you're doing test selection, you need to ask some important questions. Should you have a language-based test? What kind of physical handicaps are there? What kind of behavioral problems may be there? What is the range of cognitive level in the disease in question? And the floor --

are the floor and ceiling of the test appropriate? Will the test cover the entire range of ages and ability levels?

And you need to use the same test over and over again. You can't switch from one test to another and get any kind of meaningful data. The test must be repeatable and, you know, you can -- in young children, you need to repeat them often because young children -- practice effects are not an issue with young children because there are -- the actual test items change so quickly. And are there international -- have the tests been translated and normed? Batteries have to be short and focused, they must be relevant, they must have the appropriate metric, they must be the appropriate difficulty level, and what kind of quality control is there.

I think Delaney's going to be talking about this a lot this afternoon. It's important to compare them to other measures, for example quantitative MRI's, or ratings of disease progression, biomarkers, and all of these things are important to do within a natural history study. You also must attend to sensory motor problems, behavioral difficulties, fatigue and illness, and random variability. All of these things are important and will be addressed later today. One of the questions is, how do you know the testing is valid? And one of the things is that you use both examiner or parent to say okay,

this child is not performing well enough. You get an independent parent report and I'm going to show you now what I mentioned before, the very tight correlation that you can find between a cognitive measure and a parent report measure to give you an idea about whether the test at that moment is a valid test that you're giving to the child, who may be not cooperating or so forth. So you can sort of monitor that as you go along. I think I'm going to leave this for Kate this afternoon, but there are a lot of important factors about the testers and quality control that need to be in place before you do a natural history study that is going to be precise in its outcome.

This is the most important thing I want to mention, and that has to do with cooperation with parents in a natural history study. They're not getting treatment and, you know, parents will cooperate if you reach out to them, if you have a coordinator, like we had, who knows every patient, reaches out to them, spends time with them, explains why we need the information, and what they can get for it. And explaining the details to the parents and the child, if the child can understand, and including the parent in the testing situation, and providing feedback in a natural history study, which you cannot do in a clinical trial but you can do in a natural

history study, will allow parents to make use of the information that you are collecting.

So, in summary, measures of cognitive ability in young children are reliable and valid if correctly selected and used and with the proper metric. They should be focused and short with available normative data to make sure that you're getting precise information. In older, less impaired children, tests need to be repeatable, focused, consistent across ages, and adaptable for handicap. And correct uses of these measures require attention to quality control and examiner training. I will also mention that test should not be culture specific and we've been very, very careful to choose test that are mainly non-verbal that children can use -- that can be used in other -- even within the United States within different socio-economic groups and subcultures in the United States. Behavioral changes can also be an endpoint and I will talk about that a little bit later today. They are very difficult to quantify, but possible as parent reported outcomes. So, that's going to be something that is a challenge to psychologists.

And then, finally, parents should feel that the effort is worthwhile. This is -- all of the information I presented today would not be possible without the incredible support of the incredible team at the University of Minnesota. Thank you.

[applause]

DR. MULBERG: That was wonderful. Thank you, Dr. Shapiro, for that great, great talk.

Our next session will be comprised of three speakers. The first will be discussing the Urea Cycle Consortium, lessons learned from natural history case studies. Susan Waisbren -- Dr. Susan Waisbren is from Boston Children’s Hospital.

NATURAL HISTORY CASES STUDIES AND LESSONS LEARNED: UREA CYCLE CONSORTIUM

DR. WAISBREN: Thank you. Well, today I'm going to talk about the Urea Cycle Consortium. And this natural history is important because of the many decisions that were correctly made, because of the mistakes that continue to plague its progress, and the wisdom of the team to recognize the difference between them. And before I go further, I wanted to say that I will be focusing on some of the challenges, but this is an amazing study that in the last NIH application, got a perfect score. So, with that context, I will continue. Which one do I press to go down? There we go.

Okay, the Urea Cycle Disorders Consortium, or UCDC, is part of the Rare Diseases Clinical Research Network. There are now over 14 different centers, including two in Europe and one in Canada, that form the consortium. I receive support from the NIH and the foundations listed here for the work I do in this study and I also receive support from the NIH for some other studies and do some consulting for BioMarin with regards to treatments for PKU. I would like to thank the FDA for allowing me to participate in this workshop. I have served as the psychologist for the genetics and metabolism program even longer

than Elsa at Boston Children's Hospital for the past 38 years. I currently serve as lead psychologist for the Urea Consortium's longitudinal study and as PAI for the New England region.

To give some background, the urea cycle disorders, or UCDs, are inherited disorders that interfere with the hepatic ammonia detoxification pathway, leading to hyperammonemia and other biochemical abnormalities. Essentially, children with these conditions are at risk for high ammonia, which is toxic to the brain. Ornithine transcarbamylase, or OTC, deficiency is the most common of the UCDs and is X-linked. While most boys are severely affected or die very early, many females remain asymptomatic, although there may be subtle effects from the condition. The disorders of the urea cycle are named for the enzymes that are lacking or diminished and are shown on this slide.

But today, I'm going to focus on the four main challenges that we face in the longitudinal study with regards to neuropsychological outcomes. These include, designing and collecting data in diverse populations capturing data in uniform method in a uniform way, conducting a longitudinal study when novel therapies are introduced along the way and incorporating appropriate measures when there's a broad range of outcomes. Many of the things you've heard about today.

So challenge one: We have a neurological psychological testing data on 562 different individuals with eight different UCDs plus a few with an unconfirmed UCD. SO, this is a fairly large group. But the numbers of subjects in each group varies dramatically, ranging from two with citron deficiency to 327 with OTC. Our diverse population is further complicated by factors, such as sex, where we had many more evaluations performed on females instead of males and age, which ranged from 6 months to 71 years. Our subjects were categorized by "asymptomatic" or being "symptomatic," which meant, usually having had a hyperammonemic episode or significant developmental delay or cognitive deterioration.

The timing of symptoms and method of identification also varied. Just over 100 were identified via newborn screening, which for the urea cycle disorders only began, at most, five or six years ago and close to 200 were identified because of another symptomatic sibling, or because of a younger sibling that was picked up by newborn screening. And many were identified, these were the asymptomatic OTC females because they had a child born with OTC and then discovered that the mother had OTC. The vast majority came to attention because of clinical symptoms so already you can see the diversity here.

And then different treatments, diversities of treatments, and changes in treatments have occurred over the years, and that's the second challenge. This study has been going on for 10 years. Initiation of treatment early in life may have a huge impact on outcome, no matter what treatment is proscribed. Early Intervention and special education services may also influence outcomes and as you can see we had variability there. And there's liver transplantation that has occurred in 73 of the cases. And then there are medications. Over 100 medications have been listed in the dataset and these include medications for the urea cycle defects but also medications for depression, anxiety, ADHD, high blood pressure, or other health problems and that further complicates the picture.

In collaborative longitudinal natural history studies, one of the most formidable challenges is data capture. The initial neural psychological testing battery, which was designed over 10 years ago included over 47 different neural psychological tests, varying in six different age cohorts. This made is very, very difficult to compare performance as the children got older or to compare younger children to older children since the tests used were so different. More over, the initial test battery lasted nearly four hours so last year we

shorted the protocol to one to one and half hour sessions with the total of 21 different measures. That sounds like a lot, but when you include the parent questionnaire and the differing neural cognitive testing that need to be given to infants verses preschoolers and then we did school age and adults. It's about five per age group.

We increased the frequency of testing to every two years. Previously, if a child missed the eight-year evaluation time point, for example, he or she was not tested until 15 years of age, which was an 11-year age span. We also wanted to be sure to have longitudinal data on everyone, including adults, so, we now obtain a self-report or informant report on adults every two years as well as the every-two-year time period for the child. Finally, the old protocol did not reliably collect pre- and post-liver-transplant data, since that was very rare when the study began. So we corrected this as well.

In addition to the challenge of creating the proper testing battery, there's the challenge of actually collecting the data. For some tests, such as the language tests, which was originally administered called The Self, we only had a total of 32 testing results. For others, such as one of the parent questionnaires, the ABAS that we used, we had over 500 evaluations. Often, the sites had to rely on a rotating group

of psychologists to administer the test battery. This battery was set up, often, that insurance would have to pay for the testing or there would be just rotating psychologists from the Psychology Department of the various sites. Many of these psychologists were interns and most were unfamiliar with UCDs, possibly because the initial test battery was so long or because evaluators were inconsistent or because so many years elapsed between testing sessions or because the study coordinators thought that the neural psychological testing was optional at some of the senders. Only 40 percent of subjects received more than one evaluation. Finally, we failed to obtain information that is accessible in the database on autism, anxiety, and depression. With the new test protocol these failures are being corrected.

I've added this slide because there are two instruments I believe can be helpful when longitudinal data collection is difficult and when data on emotional well-being are needed. The first is the Adaptive Behavior Assessment System, second edition, which is similar to the old Vineland but is not quiet as geared toward low functioning individuals. And this spans the entire lifespan, from early infancy through adulthood. There are self-report measures for the adults and parent forms for the children. The ABAS, as I said, is parent

or self-report and has a broad range of skill areas related to development, behavior, and cognitive abilities; it includes 10 sub-scales, ranging from things like communications to self-care. There's four composite scores that are derived that we focused on the general adaptive composite, that provides an over all summery score. And as you can see from the slide, it correlates very closely with IQ or DQ and we recently conducted a validation study that has been published in the JIMD reports, The Journal of Inherited Metabolic Disorders reports.

Okay, the ABAS to correctly identify children performing more than a standard deviation below the normative mean on an IQ test, 74 percent of the time. For adult, the ABAS, correctly identified those performing below the cut-off 86 percent of the time. While, perhaps not sensitive to subtle changes in cognitive performance, the ABAS 2 could reliably serve as a baseline measure for global abilities. The second instruments I want to talk about are the promise questionnaires that are now highly recommended by the NIH for use in research studies. These were developed by the NIH and now have extensive datasets available for comparisons with your disease groups. These are very brief, six to eight questions each, questionnaires that measure self-reported perceptions of functioning and emotional well-being. I believe these provide

some of the best measures for quality of life for relatively high functioning individuals. The old SF-36, which some of you may be familiar with and some of those others, just didn't address the issues in urea cycle disorders or many of the other metabolic disorders that I work with. We'll be incorporating the adult questionnaires now for anxiety, depression, satisfaction with social roles and activities, emotional or behavior discontrol, and cognitive function. I list those to kind of give you an idea of the types of measures. There's other ones. There's short forms. There's also pediatric forms, which are very good as well.

The fourth challenge is variability of outcomes. I've presented here the mean scores of the developmental quotient obtained from the Bailey Cognitive Composite and IQ obtained from the age appropriate Wechsler IQ test. What I would like to point out is that among the subjects with OTC, the means scores improved from childhood -- from infancy to later, while scores from the other disorders decreased, or some stayed relatively the same. So it looks like combining disease groups may not be a good idea, similar to what else I discovered. However, we have to remember that our citron deficiency HHH nags groups only had a few subjects each. Although, I'm not showing the data, there were also differences in outcomes, depending on those

factors that I mentioned before, including prenatal verses postnatal presentations and method of identification.

So that brings us to the crux of the problem. How can we use data from our longitudinal study to select appropriate outcomes for clinical trials? Let's take arginine deficiency as an example. First, we would look at instruments, for which there's a broad range of scores. These tests here were the best from among the original tests and all were included in our shortened battery. I've told you about the ABAS, the CBC -- the CBCL is the Child Behavior Checklist. We use the WASI, the Wechsler Abbreviated Scale of Intelligence for assessments of school aged children and adults. The VMI is the Visual Motor Integration test, and the CVLT is the California Verbal Learning Test. The pegboard is a fine motor task, which may be very useful but has several different scoring methods, which led to different types of data being entered before we realized that some psychologists were entering one method and others another.

Second, we can look for data revealing change over time. But among our 20 arginine deficiency cases, only three subjects had evaluations with the same tests on two occasions, and as you can see, one case -- case one, on the first line there, improved a little. The second case improved in IQ but declined in visual motor skills. And a third case only had a

performance IQ rather than a verbal IQ because the child was nonverbal, which improved, and then what looks like a rather unreliable self-reported gain in adaptive behavior. We could also examine the impact of the number of hyperammonemic episodes, or newborn screening, or liver transplant. But with each subdivision or added variable our power diminishes and results are much more difficult to interpret.

So what are the lessons learned about developing natural history studies? First, they need to be designed for research and not for clinical care. Our initial very long diverse protocol, which was developed by a psychologist who wasn't familiar with metabolic disorders but was a clinical person and the battery looked like what a neural psychologist might have on hand when diagnosing a learning disability or identifying needed services. Second, be sure to include interdisciplinary sections. The coordinators and physicians all need to be on board and recognize the importance of consistent data collection. The dataset needs to be accessible. We might have recognized the lack of follow up or the mistakes in scoring sooner if there had been a more regular review system in place, not only when analysis were requested. Now, many of these large collaborative studies have an outside data repository and it's actually very difficult to get that data on a regular basis.

Collaborative studies are cumbersome, just know that. No matter what practices are set in place, large collaborative studies are cumbersome due to changing staff. And I know this from a number of studies I've been involved in. They have diverse populations depending where the people come from, there's inconsistent treatment strategies, even among metabolic doctors, and there are communication challenges. Don't rule out home visits in order to ensure collection of data. We've instituted that, and that has dramatically improved our data collection and we're going to be far more than 40 percent within the next year. Assess for autism spectrum disorders, anxiety, and depression, which can have a huge impact on outcomes and we didn't do that in the beginning.

So now, what do I recommend in terms of measures? Choose measures with a broad age range so you can compare different age groups. Go with more frequent testing rather than long evaluations, capture critical periods. That's something I haven't talked about, but those first five years of life are incredibly important and, as Elsa said, even those first two years, to do frequent testing then. Assess processing speed. Some people are going to talk about some of the measures to be used and then don't be cheap about paying for your psychological evaluations. Estimate about $1,000 per, and you'll get better

quality and you'll get more committed people. And I think that's often a mistake. That's kind of an afterthought in designing these clinical trials. Make sure that all instruments and questionnaires are validated. You'll be hearing much more about that.

And then, probably most important, natural history longitudinal studies are not a substitute for pilot studies prior to a clinical trial. That is why I focused on the challenges. In order to select an instrument for a clinical trial, a pilot must be take place on a group of subjects who have the same age, gender, level of baseline functioning, treatments, history, and other characteristics as a target population. I'm going to repeat that. In order to select an instrument for a clinical trial, a pilot must be take place on a group of subjects who have the same characteristics as the target population. Trying a few instruments during phase I or II, which many of the industries say "Oh, we'll try that during our phase III, and then we'll know" can be disastrous because, invariably, the instruments won't be exactly right. Then phase III is beginning and suddenly you have to choose a new instrument and it's total guesswork.

That said, the natural history study is perfect for identifying potentially useful instruments for a sub population

that can potentially benefit from a novel treatment. It can help with grouping of subjects according to specific characteristics. And finally it can high light the pitfalls that need to be avoided in a clinical trial, such as the consequences of having a very long test battery or depending on sites to find a psychologist who will follow the protocol or having an overly heterogeneous study group. I think we got to here, yup. I'm sure if you recognized the challenges you will be well on your way to a successful clinical trial, as long as you invest in pilot studies before you begin the trial so that you have it just right.

And I would like to thank the UCDs and these individuals in preparing this talk. Thank you.

DR. SHAPIRO: Our next speaker is David Davidson from Bluebird Bio.

NATURAL HISTORY CASES STUDIES AND LESSONS LEARNED: X-LINKED ADRENOLEUKODYSTROPHY

DR. DAVIDSON: I'd like to -- [coughs] excuse me -- thank the organizers for the invitation to discuss Bluebird Bio's Lenti-D gene therapy program for childhood cerebral adrenoleukodystrophy. For this talk, I'll focus on the ALD-101 natural history study, which serves as the foundation for our ongoing 102 treatment trail. Let's see. There we go. I'm a full time employee of Bluebird Bio. I've been conducting clinical research in an industry setting for almost 20 years, and I won't bore you with the rest of my bio.

Let me give you a brief overview of the disease so we're all on the same page. This is an ultra-orphan X-linked monogenic neurologic disorder that's due to a mutated ABCD1 peroxisomal transporter and this results in toxic build up of very long chain fatty acids that result in cerebral inflammation and demyelination. Now, in untreated boys, this leads to catastrophic outcomes with progressive neurologic decline and ultimately death. The only currently successful therapy is allogeneic stem cell transplant, and unfortunately, today, about 50 percent of boys are diagnosed too late to be illegible for this procedure.

The incidence of ALD is about one in 20,000 males, and CCALD represents the most severe phenotype of this disease, corresponding to about 30 to 40 percent of cases. So the average age of onset is about 8 years. The typical manifestations of the disease include progressive behavioral and cognitive symptoms. It's often misdiagnosed as ADHD in its early stages. Vision, hearing, and speech impairment occur; ambulation difficulties; seizures; and importantly, MRI pathology poses clinical manifestations, and we can quantify the white matter disease using a tool called the Loes score and gadolinium enhancement on MRI is also very important. That indicates and active neural inflammatory state, a break down of the blood brain barrier, and active disease. And as you'll see, this represents a very good biomarker, if you will, of disease progression and prognosis.

So, it terms of measuring neurological impairment the tool most widely used now is a neurological function scale, developed by Dr. Jerry Raymond. And this captures the stereotypical deficits of disease that occur over time. We, in conjunction with our clinical experts, have identified a sub group of these manifestations, and we call these the major functional disabilities and you see them in orange. They're in the major functional manifestations of disease and they

represent profound disabilities in these boys and if any of these is to occur, essentially robs these boys of the ability to function independently.

So, why the excitement about gene therapy for CCALD? Well, this is based on a proof of concept study conducted in Paris by Patrick Borg and his colleges, in which they saw three out of four of the treated boys stabilize their disease progression. They had outcomes, what you might say are consistent with an allogeneic transplant outcome and there were no gene therapy related adverse effects. Now, thanks to manufacturing improvements that we've been able to introduce at Bluebird Bio leading to much more potent and pure vector, we're hopeful we'll receive even better results in our ongoing trial.

So, I'll give you an overview of our program. There are currently three clinical trials that are a part of our clinical development. Package, the ALD-101 study, which I'll talk about. This is a retrospective data collection study in untreated and allotransplant-treated boys with CCALD. The ALD-102 trial is our ongoing treatment trial of the Lenti-D gene therapy drug product. And the ALD-103 study is a recently started observational study, including both retrospective and prospective data in boys undergoing allogeneic transplant to provide contemporary data in this group.

So, why conduct the ALD-101 study? Well, it was critical for enabling us to design the current 102 treatment trial. We needed to understand the natural history of disease in untreated boys as well as transplanted boys to identify the appropriate study population to generate the inclusion, exclusion criteria, to select study end points, and to inform expectations for treatment outcomes. Because, as you can imagine, given the rarity of this disorder, as we've already heard, it's really not possible to conduct a conventional clinical program here.

So, what does a trial look like? Five centers were involved, four in the U.S. and one in France. Initially boys between the ages of three and 15 were included, although we did include some boys with younger ages. The MRI score included 15 and lower on the Loes score. Those who were untreated were typically diagnosed after 1990. Those who were transplanted were included, typically, after 2001 onward. Transplant became the standard of care around 2000 and I believe the latest patient included was around 2010. We require that follow up data be available for at least two years on all of these subjects or until death. Case selection for the data base was based on a sequential look-back approach to minimize bias and we had a pre-specified CRF and tried to collect as much data as

possible on these boys. And the data included crucial endpoint like survival, neurological function, MRI data, neuro-cognitive data, and for the transplant patients transplant relevant outcomes.

So, let's look at the patients in the untreated cohorts 72 boys were included and the transplant cohort 65. The median age was similar, 8 for both, with similar ranges as well. And the clinical presentations for disease were similar as well. You can see that signs and symptoms were the presenting manifestation in a majority of boys in both groups. So, one of the key take a-ways of this study was that gadolinium enhancement on MRI is associated with rapid progression and poor prognosis. If we look at the first column in this table, we have untreated boys were gadolinium enhancing at the time of CCALD diagnosis, and the mortality during the observation period was 80 percent and the median survival only 2.8 years. Those who were gadolinium-negative or had an unknown status had a somewhat better outcome but still quiet terrible and a longer median survival. In the allotransplant group, importantly, resolution of gadolinium enhancement was seen following transplantation. And in those who had early disease, all of the boys had resolution of gadolinium enhancement in a median time of about three months.

So, this figure is a very vivid illustration of the association of gadolinium enhancement with the rate of clinical decline in these boys. So, on the Y axis we have the NFS score and a higher number here more disability and so as the boys descend, this is, essentially, catastrophic disease. And on the X axis we have months post gadolinium positivity and the boys in red are those gadolinium enhancing. And you can see, they essentially fall off a cliff once they're gadolinium enhancement.

So, let's turn now to look now at the all-transplant cohort. The time from diagnosis to transplant was relatively rapid, four months, although there was a wide range. We had good follow up duration on these boys. Importantly, you can see the source of stem cells was, pretty much, equally divided between bone marrow and umbilical cord blood. And it's important to point out that only about 20 percent received cells from a matched-related donor. That's obviously optimal donor source. And because of the urgency of transplanting these boys, essentially 50 percent received mismatched unrelated cells for their transplant, which as you can imagine, and I'll show you, correlates with rather disappointing outcomes in many cases.

So, the good news is that allotransplant improved survival compared to no treatment. As you can see here, we're

looking at the red line representing five years following CCALD diagnosis and we have the transplant survival of 84 percent versus 36 in the untreated boys. So, a big difference there. As I alluded to, when we look at the transplant population broken out by the degree of matching, so whether they're match-related, mismatched-related, match-unrelated, or mismatched-unrelated not surprisingly, you can see when we look at two year survival related to major functional disabilities, if they developed them or not, the matched cohort was the matched donor did substantially better than those who received mismatched cells. This really reflects one of the major unmet needs out there now for this population.

So, in addition to survival, we also have seen that transplant preserves functional ability in these boys. So here, the endpoint is the major functional disability-free survival at 2 years. So, none of these catastrophic neurologic endpoints have occurred. And when we look at the untreated positive population -- this is based on the time from CCALD diagnosis -- only 24 percent were free of these catastrophic disease manifestations at 2 years. The results are better with the overall transplant population with 63 percent are free and then substantially better yet in those with early disease. When we actually look at the untreated population in terms of using

first gadolinium positivity as a baseline, we see even less favorable outcomes.

So, as you know, allotransplant is associated with complications, typically due to immune-incompatibility issues, and that's the case here as well. Overall, we see 18-and-a-half percent of the boys experience graft failure and, not surprisingly, a higher rate in those with mismatched, unrelated grafts, a full quarter of the boys. And a substantial rate of [inaudible] versus host disease as well. Now, the Lenti-D gene therapy uses autologous stem cells for the transplant and so the hope would be to avoid any of these issues of immune incompatibility with gene therapy.

And here you just see the rates of graft versus host disease based on the degree of matching of the donor and of course, not surprisingly, those with mismatch. Unrelated donors have substantial rates of graft versus host disease.

Now, unfortunately because of the retrospective nature of this trial and the clinical course of these boys, there is actually quite sparse neurocognitive data in ALD-101, which is a little disappointing for a meeting focusing on that, but that is the case here. So, out of the 72 untreated patients only 17 had neurocognitive evaluations and only six had follow-up evaluations. The results were somewhat better in the transplant

population, but not substantially better in terms of follow up data available two years after transplant. So given then sparsity of the data collection, we could not draw any meaningful conclusions on these neurocognitive analyses, but obviously there are other very important clinical parameters that we were able to draw conclusions on the basis of.

So what were some of the challenges of the 101 study? And they are certainly those that have been highlighted by the previous speakers. Due to the retrospective nature of the trial, there is a significant amount of missing data, especially since untreated boys with CCALD did not undergo intensive monitoring. Our estimate of the two-year MFD-free survival rate from ALD-101 in the untreated patients is 21 percent from the time of the first gadolinium-enhancing positive MRI, but that number includes patients who had more advanced disease than those who are eligible for the ongoing ALD-102 treatment study. Allotransplant became the standard of care around 2,000, and so, unfortunately, obtaining a more robust dataset on the natural history of untreated boys is no longer possible.

What I'd like to talk about briefly is how we use the 101 data to inform the design of the ongoing 102 trial. And so, it was used to inform the eligibility criteria. We are only enrolling subjects with early stage disease who are gadolinium-

enhancing because, as you saw, we could predict those subjects who are likely to have rapid progression, and those who don't have a sibling HLA-matched donor, because those subjects do quite with transplant -- allotransplant. The primary endpoint is the proportion of subjects who are [unintelligible]-free at two years post Lenti-D drug product infusion and our success criteria will be informed by results we saw in the untreated population, certainly given the caveats I've described.

So, this is an overview of the 102 trials also known as the Starbeam Study, and just briefly, it's an open label multi center single arm global study. We are enrolling 15 subjects ages less than or equal to 17; they are gadolinium-enhancing; they are less for the white matter disease score ranges from 0.5-9; and neurologically they are relatively intact, or fully intact, with an NFS of less than or equal to one. The first patient was enrolled in the study in October 2013. Multiple sites are currently enrolling and we anticipate this study will complete enrollment this year.

So just briefly, some of the key secondary endpoints informed by the 101 study: obviously, the extent of white matter disease using a Loes score; looking at gad enhancement; and the resolution of gad-enhanced post-treatment; functional impairment based on the NFS score; and then survival parameters as well.

In terms of exploratory endpoints, we are looking at biomarkers and also cognitive and neuropsychological functions.

So what are the conclusions of the 101 trials? Gadolinium enhancement is associated with rapid disease progression in untreated patients; patients treated with transplant had improved survival and function compared to untreated boys; the transplant with mismatched unrelated donors is associated with higher rates of morbidity and mortality; and among patients with early disease, rapid resolution of gadolinium enhancement post-transplant was observed and maintained.

So what are the eminent needs that remain for this population? Well, we hope to reduce the complications of allotransplant with an autologous approach. And as I mentioned before, regrettably 50 percent of the boys today are diagnosed too late to be eligible for transplant so it’s really crucial to support the movement towards introducing newborn screening across the county.

So let me end by acknowledging all of those who took part in this program, our inspiring investigators and the patients themselves and the tireless team at Bluebird Bio and Veristat, for making all of this possible, so thank you.

[applause]

DR. MULBERG: Our last speaker before our break will be Dr. Anne Barbier of Shire speaking about Sanfilippo Types A and B.

NATURAL HISTORY CASES STUDIES AND LESSONS LEARNED:

SANFILIPPO A & B

DR. BARBIER: Thank you. I am a full-time employee of Shire Pharmaceuticals. My background is in internal medicine, but somehow in my pharmaceutical careers I have always worked on diseases that have some kind of neurocognitive aspect, whether they are very prevalent like Alzheimer's or medium rare like multiple sclerosis or extremely rare like mucopolysaccharidosis.

I want to talk to you about not just how we do natural histories but also what we do with the information we get from it. So what did we do before we started? What is the purpose? What happens during the natural history study? The implantation, and then what have we learned from it? I also want to explain a couple of terms that I tend to use interchangeably. Mucopolysaccharidosis Type IIIA or B is too much of a mouthful, so it's MPS IIIA or B, or also Sanfilippo A or Sanfilippo B, and PS II mucopolysaccharidosis II is Hunter syndrome. The topic is ostensibly natural history studies where no investigational treatment is offered, but ultimately, keeping the goal in mind, we want to talk about -- or we want to develop and design and succeed with interventional studies, so I will always flip back and forth. This is what we do with natural

history studies in order to do this in the interventional studies.

Mucopolysaccharidosis IIIA and PS IIIA send Sanfilippo A is a very rare disease that is caused by a single gene, a deficiency of heparan sulfate, and the clinical features of this disease as opposed to some of the MPS' is really largely neurological. Children develop normally and eventually and then people start to observe abnormalities, whether it's the parents noticing that this child has not developed as he or she should, whether it's a check-up during a well-child visit, whether it's a preschool teacher; these are the first symptoms that are typically seen. The survival can be up to the late teens, early 20s, but often at the end of a very long road of progressive cognitive decline.

However, all of these descriptions don't really capture the fact that as most of these diseases, there is a large, wide variety of presentations, as this picture shows. Some of these children have the so-called typical features of MPS. Is there a pointer? Here. And some are totally indistinguishable from a control population at first view.

So what did we have in mind when we worked on these natural history studies? Well, first of all, the obvious gain, understanding the natural history. But what does that really

mean? What is the cognitive decline? And I agree with Elsa Shapiro. It's about the rate over time. Because it's not just how much, it's also how much in one given time aspect. Why is that? Well, it will affect, immediately and directly, the duration of the clinical trial, and in theory, you can do a three-year or a 10-year clinical trial. In reality, the probability of success with such a trial is much different from that that can be done in six months or 12 months, because of changes in standard care which will affect the outcome; clinical trial fatigue and patient drop out; and also personal turnover, which will again introduce a new variable in the tightness of the measurement; and also the variability.

We tend to think in how much is the decline, but it's also what's the variability around the decline, because that is probably the parameter that has the single most importance -- influence on the size of the trial. Size of trials is a major concern here because you cannot design a trial that needs more patients than there are in this [unintelligible] disease so variability is something that we want to first of all know about, and then second, how to control.

Brain imaging. There are abnormalities and there are abnormalities. The plaques that are seen in multiple sclerosis does not mean the same thing as what you see in -- I don't know

-- cortical atrophy in Alzheimer's or what we just heard about gadolinium enhancement in adrenoleukodystrophy. Are these changes quantifiable or just descriptive? And if so, do they change over the duration of the natural history and is there a correlation with other aspects of the disease? Because there are diseases where you see all sorts of interesting brain things but doesn't really correspond to anything clinical.

Biomarkers. Are there disease relevant biomarkers and, same thing, are they stable? Or do they change with increased duration of the disease? Or are they like -- I don't know -- cortisol, which goes up and down all day. All of these have different potential applications, strengths, weaknesses when you look at them as an outcome, if not necessarily the primary outcome in an interventional trial.

And then, finally, you keep an open mind. What did you not think of but do you find in a natural history study? For instance, there are aspects that you really didn't design for but that you pick up. For instance, in mucopolysaccharidosis, these hearing issues, are they -- how prevalent are they? How bad are they? At which age do they start to occur? And most importantly, are they correctable? There are many forms, for instance, of vision impairment that can be helped with spectacles; same thing with hearing

impairment. Can you fix them with hearing aids or is this truly something that can interfere with your cognitive assessments. Feeding tubes can lead to a change of the flora of the skin which then could ultimately represent a different predilection or predisposition for infections once you start doing interventional trial that you may just as well know about up front.

The other goal is essentially test drive to validate the performance of the clinical cognitive tools that you may use to evaluate endpoints in the trial. So these cognitive tasks, they have to be standardized for homogeneous applications and drug trials, as other people have pointed out. This is not a clinical assessment. This is an assessment for an endpoint in a trial. It’s a different -- it looks the same but it’s quite different. Can you standardize this? Can you do you this training? What’s the sensitivity? If you hear from all of the parents that their children are declining or staying stable, or improving, or whatever it may be, and your test numbers absolutely do not show, this then there is probably something wrong with the test. And as Dr. Shapiro has pointed out, whenever you see that those two measures correlate with each other suggests the violent scale which is apparent reported outcome with the cognitive assessment, which is done by

psychometric specialist, and the two correlate very well you are happy because you know that they probably have some cross validity.

For MRI, going back to the question of decreasing variability -- finding a central reader, how does that work? Does it decrease the variability? How does it work to -- technical issues? How does it work to transmit those scans over internet portals? These are big files. It's not always as easy as you would think. And then for biomarkers, again, how do the assays perform? Are they sensitive enough? Are they reliable enough or producible enough? Will you end up with a shortage of critical reagent that will stop you in your tracks? Test driving all of these aspects.

The third one is a potential goal. [unintelligible] It could potentially, in some specific cases, function as a non-concurrent control. Dr. Goldsmith has explained how difficult this sometimes can be. Once there is a drug for a rare and devastating disease nobody wants to end up in the control group. Nobody wants a control group. You have to have a control group. You cannot demonstrate the value of your trial without an appropriate control group and the best control group is always one that is identified at the same time, randomized, and is subject to the same standard of care, not a control group that

is 10 or 15 years older and was described in a paper when antibiotic treatment was different, there was no stem cell therapy, whatever it is that is applicable to that particular disease. So I think this will work only if the natural history study is really matched very closely to the treatment arms of the investigational trial, so the patient population should be the same: age, gender, general status of severity. The duration of the natural history study should at least cover the duration of your interventional trials, and they should be assessed by the same tools. If you have natural history study where all the children have been assessed with the Wechsler, can that be a control arm for a trial where the endpoint tool is the Mullen? Not without some cross validation I think.

So that brings us to an actual example, the natural history study in Sanfilippo A. We enrolled children who had a calendar age of 1 year or older and a development age of 1 year or older, estimated using defined adaptive behavior scales a parent reported or a parent assessment. We did various assessments at 0, 6, 12, and 2 years -- developmental assessments, imaging, and biomarkers. We enrolled 25 children at the University of Minnesota and, very interestingly, we had six sibling pairs. And I will show you just two or three slides. I think this is essentially the same slide as what Elsa

Shapiro presented earlier, so if a child develops normally there should be perfect correlation with the development -- the calendar age on the lower axis and the age equivalent, and that would follow the blue curve. And what you can see is that patients probably start up -- or start of originally on that curve, and then they plateau, and then they go down. The slide also shows you two colors; black and red. And that seems to be two different populations, which is again important. Do you want both populations in your end interventional trial or only one of the two? We also looked at the levels of glycosaminoglycans, which are the immediate metabolic intermediate of this missing enzyme in this cerebral spinal fluid, and we found the following things. First of all, the levels in these patients are clearly elevated compared to healthy controls in green at the bottom of the slide -- no problem whatsoever in seeing that there is a disease-specific value to these markers. We also found that they are actually very stable over six months up to even 12 years. This is good, because it means that if in an interventional trial you see these markers go up, it is probably a safe assumption that this is due to the treatment since otherwise this would be stable. On the other hand, the more attenuated and the more severe patients are overlapping quite a bit. Within that population

the separation between more severe or rapid progressors and slow progressors is not as clear.

So what have we learned from this particular study? There are two different trajectories in these patients; rapid progressors -- slower progressions. The early intervention, defined here would probably be for the age of 3 to 4 years, is most likely to be successful. This is a neurodegenerative disease. Cells die so you have to probably be there before an irreversible amount of damage has been incurred and these levels of the biological marker are higher than in control populations and are stable for at least one year.

Very briefly, I'm going to touch upon some other natural history studies that we have done. In Sanfilippo B the setup is very similar. It's completed but we are right in the middle of analysis so I can't tell you yet what's going on there. In Hunter, our MPS II, we had a different approach because we were already in clinical interventional trials and found a couple of interesting facts. First of all, for Sanfilippo B we found that the speed, or the ease, or lack thereof, with which you could enroll in a natural history study, is actually a very good reality check to what you can find in the literature. Epidemiological literature in these diseases is often very scarce and very old, and can sometimes be biased by

hot spots. It just takes one big pedigree in one country to make the numbers up for that particular country. If on the other hand you cannot find patients for your natural history study, or vice versa, you are surprised by how many people sign up for this. That is a reality check for what the literature may tell you. For the Hunter syndrome, we found that if you have a natural history study and you open an interventional trial all the patients who can migrate will, and this is exactly what happened in our study. I think at least a third if not almost half of the patients in the natural history study immediately moved over to the interventional trial, of course.

So what have we learned in general? In natural history studies, sharing the results of the cognitive assessments with the families can mean team motivation. And what we have learned is that it's not just about the numbers. It's not just about your child's IQ is so much, or your child's age equivalent so much. It is the access to these highly specialized psychometric experts that they cannot access in their daily lives or even in there average -- or their daily medical environment; they ask questions that may not be related to the natural history studying; parenting advice; questions about developmentally appropriate tools and activities; deciphering the lingo that school boards use; support in

figuring out how to get to an individualized educational plan. Those are benefits that the families experience from natural history studies.

We also found out that this directly influences how to select the sites. You typically select a clinical trial site for an interventional trial based on the qualifications of the medical investigator because it's all about safety ultimately. Will this investigator be able to manage the safety of the patient? If your primary endpoint is a cognitive test you need to have a cognitive assessor who is just as good and we have had to not select clinical sites where we really like the investigator but the psychometric support just was not up to the standards we needed. A research assessment is different than a clinical assessment. It needs to be a little bit more standardized, unfortunately. In a clinical assessment it's all about coaxing the very best and most complete information out of this patient in order to get support services, individualized educational plan, et cetera. Here, it is to assess an intervention. It's more analytical. There is not immediately a therapeutic benefit or a therapeutic application in terms of, should this child need physical therapy? Should this child need speech therapy? Should this child need behavioral therapy? It is about analyzing an effect.

Much has been said about behavioral interference with cognitive testing. One of the major lessons learned is that you really do need experienced curators. This is not a task you can give to anybody who comes out of college with a degree in psychology. Here is the testing manual, go do it. It doesn't work that way. We have found that we need experienced curators. The timing of the assessment is crucial. Don't do this when the child has just come off the plane, after three hour plane ride and was stuck in traffic for three hours on the way to the clinical site, so we have the families come in the night before. Breaks. Snacks. I don't function when I'm hungry so why would the children? What we have also found is that it's almost better to have no data than to find yourself with data which has been recorded that makes no sense -- that is uninterpretable. So we have instructed the sites to gently persist as much as you can. If you can make it happen, make it happen, especially when it's a screening visit and families really want to be in the trial, but don't insist. Don't let it go from coaxing the best performance into driving it through at any cost. If it is really not possible to do the assessment that particular day, stop. Don't persist with numbers, ending up with numbers that make no sense, which then again means, in terms of schedule logistics, that we try to have always two days. If it doesn't

work -- if you start on Monday and it doesn't work, try again on Tuesday. We can build in some flexibility. Of course you cannot do this indefinitely. You cannot try on Wednesday and Thursday and Friday until you get the number you want. So moderate flexibility.

Some people have expressed concern about repeat testing and what we have heard from our experts, and which has been confirmed today, is that it really is not that much of an issue with these children. First of all, if they are young. Second, they are often cognitively impaired and the memory and retention is really not of an issue. But also, most of these tests are set up for this. They have alternative versions and the assessors note, you know, version E was given in January and when they come back six months later they give version B, etc. Psychotropic medications -- well we have to live with it. It's a fact of life. Many of these children take some type of medication. Whether it's a medication for the attention deficit type of aspects, whether it's antipsychotics risperidone, antiepileptic, which also have influence on cognitive performance, we will not ask the families to stop taking these important medications. The best we can do, as was said in another presentation, is collect the data and note this as a concomitant medication. The only thing we have asked is that,

to avoid as much as possible experimentation with doses of these medications in the trials, but still -- if the child grows up -- becomes heavier and needs a dose adjustment -- that takes precedence. We also like to make sure that these tests are not too widely available outside of the clinical trial setting so that there can be no teaching of the tests, especially when it comes to screening assessments. That very important decision about whether a child is eligible for participation in interventional trial. We don't want the test to be available off the internet so the child can be coached in this.

Another point that we've had to clarify a couple of things -- a couple of times is that participation in natural history study does not guarantee you a slot in the interventional trial if and when it comes about. Totally different inclusion criteria and typically natural history studies are much more -- have wider inclusion criteria. But the reverse is not true either; if you have been in a natural history study -- that does not preclude you from participating into the interventional study, and as I explained, for Hunter trial that has been exactly our experience. Patients who were in natural history study, and who we have very interestingly, some cognitive longitudinal data prior to treatment, and then they will go on to treatment are not depending on the

[unintelligible] and we should be able to follow them for another 12 months, allowing us to have both, essentially to have these patients function as their own historical control.

I want to acknowledge so many people who have been involved in this: the patients and their families first, the site investigators and their staffs, the psychometric staff at the sites for many of whom were very excited to participate in this but also had a bit of a culture shock when they realized you are not just doing the same thing you were already doing for this patient -- it's slightly different; it's a research study -- Kay Delaney, Dr. Shapiro, and Dr. Stein for helpful discussion in understanding all of these, and all the personnel at Shire who participated in these studies. Thank you.

[applause]

DR. MULBERG: Well, thank you, Dr. Barbier. It's now 10:20, which we just finished our 15-minute break, so congratulations.

[laughter]

No, we won't do that to you. Dr. Buracchio has informed me that we will have a 10-minute break and we will start in 10 minutes for Session 2, and then we will get into a panel discussion after the next few talks. Thank you.

[break]

SESSION II: WHAT IS EFFICACY?

DEFINING A CLINICALLY MEANINGFUL CHANGE

DR. SHAPIRO: Okay, we are going to start off this next session with some perspectives from a caregiver/patient perspective and from a clinician on what efficacy on a neurocognitive outcomes means to them. We're going to start off with Melissa Hogan who is the mother of a patient with Hunter syndrome and she's got a lot of interest in this particular area of cognitive outcomes. So Melissa, would you join us?

PATIENT/CAREGIVER PERSPECTIVE

DR. HOGAN: First of all, I want to thank the FDA for inviting me to speak today. Just a quick disclosure, my son is participated in a clinical trial with Shire and I perceived expenses related to that for the last four and half years. A little bit about me: Most importantly, I'm a parent to three sons, one of which was diagnosed at 2 with Hunter syndrome or mucopolysaccharidosis 2. In my prior life, I was a health care lawyer, but now I do a lot of work in advocacy in rare disease, Hunter syndrome specifically, also serving as an FDA patient representative on advisory boards and have a nonprofit related to Hunter syndrome research.

So I want to talk today about what's important to patients and their families related to neurocognitive assessments, specifically in the measure of efficacy, the measurement of particular domains and the measurement process itself. And also, what really threatens the effort to improve this process and make it better for both the sponsors, the FDA and families.

So we can't talk about this without an understanding of the true reality that families live in prior to an interventional trial. And what that is in a lot of these

neurodegenerative diseases is that they're watching their child crumble right in front of them and in many ways watching their life crumble. So prior to going forward for these trials and neurocognitive assessments, this is what we're looking at. So if you can imagine that, you can understand that the first and foremost goal of many parents is to save their child. So of course, we want to help the community and bring a drug to market, but their first and primary goal is to save the life of their child.

So with that in mind, what's important to families? So of course, you know, we would all love for there to be a cure for their disease, but we understand that the path is long to get there, and so, you know, improvement would be great, but really, stability would be fantastic. And frankly, because we see the science progressing so rapidly since it has come in the last 10 to 20 years, we think disease progression -- slowing disease progression is still pretty fantastic because we know that other things are coming down the line.

And what we call those things is our inter-cure. This is our cure for now until something better comes along, and a lot of us are totally fine with that. I think we sat in this very room for the patient focus drug development meeting and a parent of two children with Batten disease said we just want a

lifeboat. So we're not looking for the best therapy; we're just looking for a therapy to last for now.

Sorry. So what does that look like? Stability is much better measured by, as Dr. Shapiro mentioned, age equivalency or achievement scores versus a cognitive score, because if you understand even children who are stable, who are in that range of 3 to 8 years old, will still show a declining cognitive score even if they're stable. So age norming is a problem in that area. So in the measurement of domains, what is important to families?

So behavior, we've talked about that a little bit already. How much of an impediment that could be to neurocognitive testing but just in the measuring the efficacy of a drug, behavior is an important aspect of families because it really impacts their day-to-day life with their child. You see in this picture, before intervention my son was in a pediatric wheelchair with a six-point harness because he had no, absolutely no safety awareness so that if you can imagine that very much impacts your day-to-day life and the function of your family. Activities of daily living -- so independence, are they able to feed themselves, are they able to toilet, are they able to help get dressed, tell you if they're in pain, tell you

anything, have speech at all. So those things are hugely important to families.

And finally, cognitive ability is really the last thing that's important to families after those first two and really, it's only important to the extent that it's a proxy for the disease progression. So when you're told your child is going to die, you could really care less if they learn to read. So their ability to function in the family and manage behavior and manage independence are the most important things. So and why is that important? Because that makes the importance of parent reported measures or child directed or observational measures much more important than standardized neurocognitive assessments that are normally given by an assessor.

So, in the measurement process, what's important to families? First of all, I would say flexibility. You need to first be flexible in order to figure out how to make this happen. How to make these visits happen and the assessments happen but after that you really need consistency. And that is consistency for the child and the family internally from visit to visit as their coming back but is also we talked about from site to site. So there examples of let's say you have a site where you have an assessor who is very cold and clinical, they're not going to get as good of results, their

neurocognitive or accurate results, as you are with an assessor who understands the disease and works with the child in an upbeat and excited manner. Or for example in your environmental aspect. If you have a room for neurocognitive assessments that is white and is boring for children with an attention deficit problem that is much better than an assessment room that has a bunch of stuffed animals and a computer and lots of interesting things and then also laterally, patient to patient. The consistency is needed there because, for families to endure these long natural history studies or trials, they depend upon one another and part of that is knowing what to expect and how to handle that logistically, especially if you're traveling on a month-to-month basis to a site for long periods of time.

Also what's important is value to these families, and I think we talked about that in terms of natural history studies in giving back information and what's helpful to families but even in interventional studies. We have to understand that these are kids and during possibly cognitive testing every three months and other testing as well, but they also have a regular life and that is in school. They may have to suffer additional testing in order to get services in school. So if you can imagine layering that on top of it, it's very challenging and so to give information back to them is incredibly valuable. And

along with that, minimizing medical trauma. So that goes along with value. The medical trauma that can result from children being involved in clinical trials and natural history studies is not just from medical interventions like, you know, pokes and blood pressure and things like that.

Neurocognitive assessments also comes into that because it really -- what medical trauma results from is really getting a child to do or putting upon them things that they don't want to do and neurocognitive assessments are part of that. So in our circumstance, my son suffered two years of very severe medical trauma that we had to work through as a result of all the things required of being part of a clinical trial and severe medical trauma can also be bad for the sponsors because it can result in a child not qualifying for a trial or a child being asked to leave a trial because they cannot function within what is required of them.

So what threatens the sufferer? So the biggest thing we've talked about a little bit is the heterogeneity and that's not just the heterogeneity we think of patients. Of course, there's heterogeneity of disease in the patients and I'll talk about that in a minute and a couple of examples. But there's also heterogeneity in the parents. There's heterogeneity in their knowledge of the disease and how they deal with it on a

day-to-day basis, how they handle their child's medical trauma, the resources they have, how they access services through school and through therapies. So we can't control a lot of the heterogeneity in the patient population or we can't. We can affect heterogeneity in the parent population with knowledge and education.

And then there's two other areas that we can control through proper quality control methods and that's the heterogeneity of the clinicians, whether it's the main clinician or the assessor for the neurocognitive assessments but also the environment of the neurocognitive assessments. So I think it's really important to think about that up front and how that's going to be handled. So, in particular, an example of heterogeneity disease, if you look at all these aspects that's just all the areas where there can be variability, just in the example of Hunter syndrome. So that's something we can't control so then it makes all the more important to try to affect or control the heterogeneity and the other aspects. Speaking in cognitive aspects, the heterogeneities, you see a very consistent line in three of these children but then you have the outlier of an older child who's more stable and actually shows a little bit of improvement so taking that into account up front and how you're going to deal with that.

So we looked at heterogeneity, also, what threatens the effort to make this the most effective for everyone. And I don't often get to quote Eminem, so I'm going to use this opportunity and be cool with my kids: "You don't get another chance; life is no Nintendo game." And what I mean by that is we can't short-sell the impact of the decisions related to neurocognitive assessments. Getting into a trial for families is often literally whether your child will live or whether they will die. So we have to understand that. So that makes it all the important -- the use of surrogate endpoints also instead of an arbitrary cognitive window for enrollment that aren't safety-based exclusions for a trial, because that results in things -- situations where a parent may save one child and then watch one child deteriorate and die. It can also result in a child having one bad day and not getting into a trial, which again is a life or death situation in a lot of cases. It can also result in children missing a trial by one point, again not a safety criteria but just, again, a particularly bad day.

And the age norming problems I talked about earlier exhibit a problem in that in such that a 3-year-old with the exact same set of skills will get in a trial, but a 6-year-old with the exact same set of skills will not get in the trial when

we have these arbitrary age-normed windows for criteria eligibility.

And finally, the commitment of families. We have to understand that that threatens the effort of participating in trials that require very extensive neurocognitive assessments and just educating them and giving value to them will help in that regard.

Finally -- I've said this a number of different times in different settings -- everyone has their specialty, and doctors know a lot about the disease, and the sponsors know a lot about the drugs and the disease itself, but really, an understanding of the day-to-day course of the disease and how the disease plays out and why that's important in neurocognitive testing and what you can measure in a better way and how the drugs work, that really comes from the patients, so it's really important to have them involved as early as possible. This is just my connection points. And thank you very much.

[applause]

DR. SHAPIRO: The next speaker will give the clinician perspective, and Jonathan Mink, M.D., Ph.D., from the University of Rochester.

CLINICIAN PERSPECTIVE

DR. MINK: Good morning and thank you and I have no slides. My name is John Mink, I'm a pediatric neurologist at the University of Rochester and I have a number of areas of interest but movement disorders, cognitive behavioral neurology and neurodegenerative disorders are kind of been my area of focus and I was asked to talk a little bit about the clinicians perspective on what are clinically meaningful outcomes.

This is an area where I've been very much influenced by some of my work in cerebral palsy where I've interacted a lot with the physical medicine and rehabilitation community, a lot with the disability community where the WHO and others have developed these different levels of looking at the impact of disease on someone's function. And all though we clinicians, we physicians, researchers, biochemists, biologists, often focus on what's called the path of physiology, what's wrong with the cells, what's wrong with the chemicals and what we use in the clinic often is what is often referred to as the impairments, the weakness, the processing difficulty, the delayed reaction times, it's really the function and the societal participation that is key. And then I'll get to quality of life, which encompasses so many of these things.

The disorders we're talking about today, for the most part, although these are biochemical disorders, they're not like diabetes, for example, where you can say you know a reduction in the hemoglobin A1c percentage from 11 to 7 is a really good thing. And although the patient's may not feel much different and may not be able to do much differently, then we know that that has tremendous impact on their risk of developing complications down the line. So we can celebrate a number. But for most, the disorders we're talking about today, we don't have that kind of number where we can say, "With that number, we can predict certain things for certain." And even the enzyme deficiencies are those where we can measure molecules that are part of intermediary metabolism for the most part we can't say that our treatments may or may not lead to a change in enzyme or metabolite, that those numbers really provide a very meaningful difference. And so what we're left with is what the patient can do.

And I spent a lot of my life using rating scales and research and I get excited over a five point change but I've never had a parent come in to my -- to me with their child and say, "So are you going to tell us today that there's been a three point improvement in that rating scale." That's not important. That may be important to the research and I think

it’s very important to have that information but that’s not what parents and patients care about. We -- most of the disorders we’re talking about today, most of the disorders that this kind of work is focused on, are progressive disorders. There are some that are more episodic; there are some, as we’ve heard about before, that can be asymptomatic for a very long period of time. But most of these are progressive disorders; most of them lead to premature death. And so our goals may be very different.

Ultimately, what we all want to do is completely reverse the effects of the disease and restore all normal function and make it no longer an issue. And we have almost no disease where we can do that today. Most of the interventions, and there’s been a huge explosion of experimental therapeutics research in these disorders over the past five to ten years but for most of them, realistically it’s stabilization, it’s halting the progression that is the most realistic. And then there is some where we may have some hope for some reversal at least around the edges.

We also think about treatments that provide symptomatic benefit. We’re pretty good at treating pain. We’re good in instances like Parkinson’s disease by giving supplemental levodopa. We can improve some of the symptoms but

that doesn't affect the underlying disease process. We're really quite good at treating epilepsy but our treatment of epilepsy in these disorders that have seizures as a component rarely has an impact on the underlying disease process. And so again as a clinician, we can often see these as victories. A child who goes from having 20 seizures a day to no seizures a day, that often is accompanied with a substantial improvement in the quality of life for both the child and the family but it doesn't really have an impact on what's going to be happening two, three, four years down the line.

And so then there are treatments that are focused on disease modification. In slowly progressive diseases is a clinician that's very hard to access. A disorder, so much of work is in Batten disease, and I'll talk about this this afternoon, but the most prevalent form of Batten disease, juvenile Batten disease, has symptoms starting around age 4 or 5 years of age and progression to death sometime between age 20 and 30 years. And so it's very hard if you're seeing a patient every four to six months to get a good sense of really how things change, and, in fact, in some ways, it's even harder for the parent who's with the child every single day for hours a day to have that broad picture of how changes are. And so sometimes, you need to step back and look at how things are

today versus three or four years ago. So those time points can be very important, but there -- it's very hard to kind of wrap your head around some of those things.

We have rating scales. We have different scales that have been used to access severity of things. You heard about some of those with ALD for example. Some of these are very well validated, very reliable. Some of the neuropsychological testing that Dr. Shapiro was talking, you know, these are well established measures and they're important but again that's something that in a clinical visit as a physician I take that into account but those rarely are the kind of end points that we're looking for when we're gauging the impact of some kind of intervention. They can also be very insensitive. One of -- I -- my life as a movement disorders neurologist, the unified Parkinson's disease rating scale was one of the first really well validated reliable wide spread used rating scales for a neurologic disorder. And it is a very good rating scale but it has 132 points, zero being no problems to being the maximum you can have and 32 of those points, if I'm not mistaken -- maybe it's only 28 -- are for tremor, and four are for gait. And when you ask people with Parkinson's disease what really limits their life, it's the falling, the freezing, the inability to walk independently, and the tremor may be an annoyance for many

people, but it doesn’t cause them much disability. But we can measure it in every limb and we can measure it in different settings, and so we can get lots of points for it. So some of the rating scales may be very, very useful, well-validated, highly reliable but not meaningful for what’s important to the patients.

And I’ll give you an example, just a patient story. So many years ago, I saw a 16-year-old girl who was a -- had cerebral palsy. She had spastic and dystonic cerebral palsy. She had this beautiful motorized wheelchair that she could not drive because her movement disorder was so severe and she had so much shaking, too, that she could not manipulate this in a safe way. And she could not hold a cup to drink from. She really could do almost nothing for herself. Her cognition was quite normal but her communication was very severely impaired. And we gave her a trial of medication and we used all the available quantitative rating tools we had and she had a one point which was a clinically insignificant improvement in her motor function measures but with this medication she could with one arm now drive her wheelchair, she could hold a glass of water with a cup with a top and a straw so she could hold a cup and drink for herself. So manipulating, moving herself through the environment with a wheelchair and being able to independently

decide when to drink from that cup was an incredibly enabling for her. We couldn't measure that by any rating scale. It was both quality of life but it was function. And so I think as a clinician, increasingly I'm trying to focus on things that are functionally meaningful and that lead to improved ability to participate in society whether that's school or going to the mall or whatever.

And then, finally -- and this is where I think those of us who are involved in both clinical research and clinical care, we often talk about translational research -- I think a lot of the things we learn from interacting with families who have family members who have rare diseases, both as a researcher and as a clinician, the research experience informs how I practice in my practice, and my experience informs how I think about the research, but quality of life is so important. And the mantra in pediatrics -- pediatric neurology, the mantra is, first of all, children are not little adults, and second of all, the quality of life of the family is something you can't ignore. And so the quality of life of the family, the ease of care, the burden of care, the anxieties associated with having a child with a progressive disorder are things that you really have to take into account and consider when you're evaluating the impact of a treatment of the child. That you may see some little tiny

areas of improvement in the child but the important thing is not being addressed and you mentioned behavior. You know, behavior is important and you may say, “Okay this child can now sit up for a little bit better” but if their anxiety and their perseveration and their aggressive behavior is no better than it was before the impact of that on family functioning on the function of the child may be much less than you might expect.

And so that’s my perspective as a clinician on outcome measures and in how to judge the effective treatment as I think that all the different things we can come up with like MRI scoring, like quantitative testing is very important to guide research and to help evaluate the efficacy of a treatment but as a clinician it’s really the impact of those treatments on function, societal participation and quality of life that are most important. Thank you.

[applause]

DR. SHAPIRO: Our next talk is by Dr. Florian Eichler from Massachusetts General on assessments in early intervention for presymptomatic disease.

ASSESSMENTS IN EARLY INTERVENTIONS FOR PRESYMPTOMATIC DISEASE

DR. EICHLER: Thanks, Dr. Shapiro, and I'm grateful to the organizers for having me here and what I perceive as probably the most challenging title that I was given, which is assessment in early intervention of presymptomatic disease. So this took me some time to wrap my head around, what this actually meant, and made me also reevaluate a lot of the things I do.

Let me just start off with disclosures. I've received [unintelligible] Alexion for consulting as well as from Retrophin in developing therapy for CTX. I'm co-PI on a gene therapy trial for ALD, sponsored by Bluebird Bio. So I'm an associate professor of neurology at Mass General and really was brought into the field by Hugo Moser at Johns Hopkins many years ago, and that has, over time, led me to start new treatments in the field of Inborn Errors of Metabolism. I'm currently PI of a trial of [unintelligible] antiviral chain correction and also have hereditary sensory neuropathy funded by the FDA, placebo controlled, randomized trial. Importantly, I also am PI of a patient-powered research network contracted by PCORI, and I think it fits very nicely into what you just heard in terms of

the importance of patient-reported outcome, and I'll briefly touch upon that.

So what is presymptomatic disease, and why should we care about this? I will briefly touch upon this and discuss the optimal timing of interventions in Inborn Errors of Metabolism and I think the understanding presymptomatic disease really fits well into this question and then lastly what are the optimal presymptomatic assessments? So I'm shifting gears here slightly from neurocognition towards this presymptomatic disease state and I think this will hopefully lead to a nice discussion afterwards.

So what is presymptomatic disease? So I opened The New York Times a few weeks ago and here in the Sunday Review there was an article about President Ronald Regan and his speaking patterns that were analyzed by a group that had recently published in The Journal of Alzheimer's disease. And they found out that years before doctors diagnosed him with Alzheimer's there was a change in his speaking patterns linked to the onset of dementia. He used repetitive words, there was substituting nonspecific terms like "thing" for specific nouns and there was a decline in the use of unique words. Now we don't think this impacted his decision-making ability but it was an interesting study that was quite rigorous about this type of

comparison and I thought was sort of probably the most publicity that was given to the presymptomatic state recently.

So why do we care about the presymptomatic disease? If we found early specific and sensitive signs, I think this would dramatically change how we treat patients. We could have an earlier window of intervention. We could treat -- tailor our treatment towards specific phenotypes because we would have a better understanding of how the disease would evolve and what phenotype would come about. We could have earlier measures that we could follow over time and ultimately more successful trials as we're not treating to late and you've heard some of our recent experience in adrenoleukodystrophy. Okay.

So what's the challenge in assessing presymptomatic disease in children? And Dr. Shapiro has already spoke eloquently to this. So I think one of the challenges I see also as father of two boys, 4 and 5, that the time of attaining milestones is quite variable and each child develops skills at a slightly different time point and that most importantly the structure of our brain is not complete. Myelination is an ongoing process and one of the challenges we have in Inborn Errors of Metabolism is differentiating that ongoing myelination from its deterioration. So really essentially differentiating developmental delay from regression, okay.

What is simply delay and what is loss of skills? And here's a picture of one of the first depictions of myelin, this insulation of the wiring in the brain by Paul Flechsig back in 1921. The take home message here is that there's very little myelin that -- around birth in the brain. And myelination really just takes off in those first years of life and it starts from the peripheral nervous system and moves up into the central nervous system and it starts from the sensory nerve fibers and moves into the motor fibers from the back of the brain into the front of the brain. So if you look at a child's brain MRI at birth there's little myelin and presymptomatic myelination status may reflect development not disease. So I think that's a very tricky situation that we have to put our head around. And you really have to understand here on these T2 weighted imaging and I won't bore you with the details of MRIs, but everything that's black and dark here on T2 weighted imaging is myelinated and what is light is not. So there's only very little myelin here in the brain stem and just a few parts of the supratentorial regions. So our success in evaluating cognition and development is much better once myelination is complete and I think those Inborn Errors of Metabolism that have completed myelination before the disease strikes are we're in a better position to assess presymptomatic disease. So [unintelligible]

is one of those where really myelination is complete and then the first lesion is developed.

So you've already heard about this monogenic disorder due to mutations in the ABCD1 leading to elevations in very large fatty acids and then a myriad of different phenotypes and I'll briefly talk about this. There's a dysfunction of the adrenal gland, there's demyelination in the brain and then there's an axonopathy of the spinal cord. So ALD is progressive but has phenotypic variability and so I'm just pointing this out here because the same situation is present from many Inborn Errors of Metabolism. You have one gene, you may have one pathway but then the expression varies and the phenotypes are diverse. So in the case of X-ALD you have the cerebral form of the disease, denomination and inflammatory response in the brain. There's a little boy who I've taken care of who saw the Pope recently, he's wheelchair bound and nonverbal and on the other hand adrenomyeloneuropathy in these patient's, their brain is not affected and they have spinal cord problem and axon degeneration in the spinal cord leading to gait, bowel and bladder problems. So this phenotypic variability needs to be understood and if you look here at these different phenotypes, you understand that on the one hand there's AMN that leads to a normal brain but spinal cord degeneration; on the other hand,

you have this severe childhood form and if you look at the brain MRI you can see how initially the brain is normal and then over the first few years of life a lesion develops and then by 10 to 15 years of age, there’s a lot of devastation.

So in Inborn Errors of Metabolism, presymptomatic assessments have long focused on biomarkers and this is because it’s quite attractive so you understand there’s a single gene that often interrupts a certain enzyme and leads to build up of certain substrates, ALPHAtides, metachromatic leukodystrophy, very long chain fatty acids and ALD and [unintelligible] disease. However, chemical biomarkers may reflect the path of physiology but do not predict the disease course and that’s the big challenge here. So even though plasma very long chain fatty acids are an excellent biomarker -- they are great for diagnosis, they are easy to measure, highly specific -- I tell all my colleagues don’t do a brain biopsy; you just can measure this in blood. But they do not predict phenotype, and that’s our big challenge.

Much better in ALD is brain MRI, and we’ve been able to show that, and this is actually the work of Hugo Moser of 20 to 30 years, that brain lesions -- scored from zero, not involved, to 34, highly involved -- correlate with survival and if you have a -- no lesion in the brain then your survival is

much better or your lesion is very small and as opposed to having a higher score, then your survival is much poorer. You see how this confluent lesion spreads out and eventually takes over all of the white matter. So brain MRI in this case in the presymptomatic phase is sensitive and specific and develops usually in the center of the corpus callosum and then as you heard there is a period of contrast enhancement followed by this rapid evolution. Importantly the symptoms don't arise until the lesion has become of a certain size. So really, you can follow this whole presymptomatic phase just by imaging alone.

So why do we care about early assessment and good predictors? I think in medicine we've done often too much and to advanced stages in the past and I think that's been a big problem. We have a lot of technology at our hands, we can use it but is it meaningful and does it improve quality of life. And I think historically we've often too little throughout the time period of a patient's life -- and here I found the first photograph of a boy with adrenoleukodystrophy back from 1910. Looks very similar to some of the boys with advanced disease. So we care about these stages because optimal timing of intervention determines the success of Inborn Errors of Metabolism. There is a strong correlation between the age of symptom onset and the severity and rate of decline and here just

the different phenotypes of metachromatic leukodystrophy with late infantile, juvenile and adult forms and you can see the later the onset, the shorter the five-year survival period. And if you treat in the advanced stage, you can clearly see that your success is going to be much worse than compared to an earlier stage. So you really want to get to that early stage that often relies on presymptomatic assessments. And to prove this point here, I want to show you some results of bone marrow transplantation. You've already heard about this. If you treat boys who have a very low lesion score, if their lesions are very small, your chance for successful bone marrow transplantation is much better and their survival is much improved compared to boys with advanced disease. So brain MRI matters and it helps determine the optimal window of intervention. And this is true also, if you combine your MRI lesion score with that of neurologic deficits and the same improvement in survival has been shown when comparing those with few neurologic deficits to those with several neurologic deficits.

So can we do better than conventional imaging? And you've heard about contrast enhancement already briefly. This contrast enhancement correlates very nicely with the lesion size evolution. If you have contrast enhancement, you're lesion size will become greater in the coming years. This is here -- the

MRI imaging score that indicates the lesion size as opposed to those boys who don't have contrast enhancement whose lesions generally remain stable. So we've recently gone further and now looked at MRI perfusion and understood that in the areas that are not yet affected, you can even hear/see low cerebral blood volume and that that is predictive of the evolution of the lesion. This is true across phenotypes and holds true both for childhood cerebral as well as adult cerebral disease. So we've used this to predict whether patients will have lesion evolution and if they have low cerebral blood volume you see that the area of contrast enhancement moves forward as opposed to after successful bone marrow transplant where their contrast enhancement goes away and the lesion stabilizes together with the perfusion.

So I don't want to say that it's the disease alone, why you have to treat early. I also think that intervention itself takes time and we have to also take that into account. Often it takes time for our bone marrow transplantation to reach its target, for the cells to move from bone marrow to blood and from blood to the brain where we think they become micronuclear [spelled phonetically]. And the same is true for even the more advanced treatments such as gene therapy where our early read outs are that of protein correction, we can measure the amount

of ABCD1 in the peripheral cells and in these early symptomatic stages we can watch a very long chain fatty acids go down. But what's really meaningful is what's happening in the brain and what is happening in terms of the structural evolution of the brain. This is the untreated condition and here in this science paper you see that the lesion halted within 24 months, really giving meaningful quality of life to these boys.

So I just want to briefly turn to metachromatic leukodystrophy and say the same holds true in other leukodystrophies and here a recent science paper showing lentiviral hematopoietic stem cell gene therapy that's benefiting early onset metachromatic leukodystrophy. Here, these children gained gross motor functional skills as opposed to the many untreated children. Some of them even improved in their nerve conduction velocity and importantly did not develop the kinds of lesions you see in the advanced untreated condition.

So what are the optimal presymptomatic assessments in neurological Inborn Errors of Metabolism? So after everything I told you, you might think I'm advocating for using MRI for everything but it's not true, okay. So optimal presymptomatic assessment depends on what part of the nervous system is affected and you really want to think about your disease and

understand the biology of your disease. So if your brain is affected in leukodystrophy, that's where you want to look. If it's on the other hand, it's the alpha motor neuron and the spinal cord then you want to use different assessments and then if its' your peripheral nerve you want to carefully look at nerve physiology. So here's a slide I got from my colleague, Kathy Swoboda, that really emphasizes the great value of nerve physiology in the presymptomatic phases of spinal muscular atrophy. Here there's little value of brain MR imaging but you want to really go where the pathology is about to occur. So here compound muscle action potential, motor unit number estimation, electrical impedance, myography, all are very sensitive in the presymptomatic stage and Dr. Swoboda has shown that really using ulnar CMAPs, you can distinguish the different types of SMA very early in the course. And this will help you predict how fast these children will decline. We've looked at Inborn Errors of Metabolism such as HSN1, where we have a wonderful biomarker but what's really helping us in terms of predicting the future course, is looking at the intraepidermal nerve fibers. So this is a peripheral neuropathy that affects really the distal limbs of arms and legs and if we do little punch biopsies in the thigh and in the distal leg, we can directly look at the nerve fibers. What a wonderful opportunity

for a neurological disease, right? And if you look at those nerve fibers on the intraepidermal space, you can see that most of these have been lost in the thigh and I have not come across a single patient in the presymptomatic stage even who has distal leg intraepidermal nerve fibers.

So let me close now with one final point about the importance of patient-reported outcomes, okay. We've, within adrenoleukodystrophy, created a consortium around patient centered research and really brought together many stake holders in the field across patients, patient advocates, scientists, physicians but also industry participants. We said, "We did not want to exclude a single group that was invested in making progress in the field." And just within two years, we've rapidly grown, we became one of the 18 patient powered research networks funded by [unintelligible]. I learned in the process a lot from my patients and fellow advocates. And so here's a chance really I think also to gain presymptomatic insights in to what's important to patients and I often find parents, you know, keep wonderful records of their children and you can learn a lot about their own social milestones that may be crucial to the presymptomatic stages. And so through patient learning academies and patient portals and websites, we've been able to take advantage of some of that information. So here's just a

page on the top of our -- from one of our patient portals, an example where we've created simple tools online where patients can use cursors to make statements about their own health. And, you know, and this goes across many different domains, walking, balance, speech, hearing, pain and mood and here's one about their future health and they can decide, I'm hopeful about my future health or I disagree I'm not hopeful about my future health. But it's really in their power to make those statements.

So in conclusion, I think chemical biomarkers and Inborn Errors of Metabolism are attractive but not necessarily predictive of disease course. Presymptomatic biomarkers that reflect structural deficits to the nervous system have a higher likelihood of being predictive of change. Meaning go to where the pathologies occur. And so you know where the crime scene is, that's where you want to put your marker, right. So structural vulnerability, the selective vulnerability in Inborn Errors of Metabolism is what determines the best assessment. Brain MRI and perfusion for cerebral ALD but for metachromatic leukodystrophy you might include nerve conduction; in SMA, you might use nerve physiology. In hereditary sensory neuropathy, you want to maybe directly go for the peripheral nerve and the small fibers in the skin. So it really depends on the disease.

And I think in the coming years, we'll really see that patient-reported outcome will enhance nerve cognitive assessments in the presymptomatic stages. So early intervention and presymptomatic disease can be transformative and I just want to here show our community that is been really quite galvanized by new born screening coming about, us having a consortium and really some treatments that are really changing the lives of patients. Thank you.

[applause]

PANEL DISCUSSION AND Q-and-A

DR. SHAPIRO: Will all the speakers please come to the front please and have a seat?

DR. MULBERG: Exciting morning of talks. I know that we're all very eager to get into the questions and answers. We really do invite the audience to please step up to the microphone. In the absence of any immediate questions, obviously Dr. Shapiro and I and -- will try and initiate the conversation but really do invite the experts that are sitting here to please come and provide us with a thought provoking issues that I know are causing you unrest.

Well, I'm going to start off, if I may, no specific order. I really enjoyed all the talks. Dr. Florian, may be a question for you which maybe dovetails on a couple, but on the patient-reported outcome, you know, aspect, I was noticing your inventory which some of the questions seem very difficult necessarily like my nerves are working well. Maybe you can just comment on what you're going to plan on doing with that patient-reported outcome inventory for your research.

DR. EICHLER: So the approach we've taken is to try to harmonize and aggregate data from various different sources. So the approach we took is to say, patient-reported outcome was

important to us but we are creating a platform where we also having physician-entered data occur at the same time. And we're using global unique identifiers to link those data sets. So yes, there might be difficulty in answering certain questions but you can validate through comparisons with physician-entered data as well. We've also found it important to phrase questions in a certain way from the patient perspective and doing that really made a big difference.

DR. SHAPIRO: What are you doing with parents -- with children who are too young to fill out something like that? Do you have a parent form of that or, how are you doing that?

DR. EICHLER: Yeah, yes. So there are -- you can on the patient portal, you can log in and create your login member and then decide on whether you're filling out a form for yourself or a form for your child, and depending on that different forms pop up and then your consent you sign is also different.

DR. SHAPIRO: And have you run validation studies on both forms or what? How are you handling that, the sensitivity and liability of those measures?

DR. EICHLER: Yes so all these things are in process. What we started off with is with the physician-entered forms, which have long been used for 20 years in the field but no

adapting them for patient entry in order to allow for those comparisons. But yes, I think they're -- we're still in the testing phase.

DR. MULBERG: I can see Dr. Parisi walking up for a question.

DR. PARISI: Hi, Melissa Parisi from NIH. Yes, I have two questions completely unrelated to each other but I'll start with one.

Susan Waisbren thank you so much for your presentation about the urea cycle disorders consortium, which is an effort that we at NICHD and other have funded. But wondering if you could speak to how you deal with sort of a midcourse correction with regard to the neuropsychological battery that you embarked upon initially when it seemed like a good idea and then realized in retrospect you needed to alter and how you can use those data longitudinally?

DR. WAISBREN: Okay. Thank you. Yes, that was a long discussion. We have a psychology network and we meet on the phone periodically and one way that the statisticians are working is to try to look at functional groups and see if they can get standard scores that can use some of that -- the data that is so missing but we do have the important data. Another way is to do sub studies looking at instruments that did have

adequate data, some we are going to lose because it just wasn't given enough times. And then there are, you know, because it's been going for 10 years, we are going to be switching to the more recent norms for the IQ tests, especially for the preschool, which is just coming out. And so again, we will make the statistical corrections in that way. But that is a challenge with longitudinal studies, that there are -- do you use the old norms or do you move up to the modern norms. And we decided to move up to the modern norms.

DR. SHAPIRO: Can I argue with that?

DR. WAISBREN: Sure.

[laughter]

DR. SHAPIRO: So, you know, we have been doing a longitudinal study also and the Rare Disease Network for MPS-1, 2, and 6 and during the time of our five-year study, and we have a very exact and easily replicated from one site to another battery, during the study the norms were -- one of the tests was re-normed and we decided to stick with the old form because we didn't want to have a situation where we couldn't compare the scores before and afterwards. And now we're in the situation where we have another five years of funding and what do we do at that point. So then at that point, we made the switch. But I think that this is a very important problem in the whole field

of neuropsychological assessment. You know, it's sort of like the test companies are out there to make a buck and so they re-norm these tests every periodically and they get a lot of money because people have to buy the new test and everybody has to keep up with it. And so, you know, this is real problem for clinical trials, and you start out with one test, you really can't change that test during the period of a clinical trial or even a natural history study when it's defined in a certain way. Now, you can make a shift when you have 10 years of data and then you can take -- do some statical corrections but it is a real challenge. John.

DR. MINK: If I could ask the psychologists on the panel to weigh in on this issue. So this is something we see clinically in child neurology where we have patients referred because they appear to be regressing because they are no longer at the expected age norm. And you talked about the using the ratio and but in a clinical trial setting in particular using norm versus raw items where a decline in a norm score may not reflect loss of function, it may reflect lack of expected improvement whereas looking the raw items. So when you're talking about degenerative diseases, do you ever use both and is it justifiable to do that?

DR. SHAPIRO: To use both like an age equivalent score and a standard score?

DR. MINK: And do you use raw scores for specific items? So a standard score is also going to be age-corrected --

DR. SHAPIRO: Yeah, yeah.

DR. MINK: -- but if you have for example someone who had a, I don't know I can't think of a specific example, but where they had -- you know, could name a certain number of items at one age, and they named the same number of items at the next age. That wasn't really a loss of function that was a loss of expected gain; where as if they named fewer items that was loss of function.

DR. SHAPIRO: That's correct. Exactly right.

Jonathan Waisbren: And do you make those distinctions in the clinical trial setting and can you?

DR. SHAPIRO: Yes. That was the point that I was making with the younger children, that you really need to do that using age equivalent scores because age equivalent scores will give you that information about the rate of change of a particular function in a child. You know, for example let's take vocabulary. We don't usually measure that but, you know, the child has 10 vocabulary words at one age and then they have 10 at the next and that's not good because they should be

acquiring those skills and so, you know, that's important data for you in a clinical trial or a natural histories study. On the other hand, you also want to know what the relationship is to the normative data. So you might want to include both kinds of information or you might want to take that raw score and have a DQ, which is a ratio of the age equivalent to the chronological age, which is what we did with the Sanfilippo trial in order to look at how those children progressed both from within their own performance and in comparison with children who are normally developing.

DR. BARBIER: And I want to chime in here again to stress the importance of an appropriate control group because you might have that child who is essentially stable in what skill he possess but whose standard score is going down because his pierce [spelled phonetically] progress and to some extent it's a confusing picture. Why is the number going down, he's doing exactly the same things he did a year ago? When you look at the effect of an intervention, you need that control group to say, well yes, his numbers go down but the control group goes down even more.

DR. SHAPIRO: Right, exactly.

DR. BARBIER: So it's the change for the individual child if the change compared to the healthy control and then it's compared to the in trial control.

DR. SHAPIRO: Yes, exactly.

DR. WAISBREN: Can I just add one more psychological viewpoint, which is that whenever possible, you want to use the standard norm scores, and what Elsa's talking about is when you have a population that really is below the floor.

DR. SHAPIRO: Right, exactly.

DR. WAISBREN: And so when we're talking about the urea cycles, where most of the children aren't below the floor at this point -- some of the adults are, but most of the kids aren't -- it's much to our advantage to use the norm scores.

DR. SHAPIRO: But then you get into this issue of the re-norming of tests and you know the justification for that is something called the Flynn affect, which is that IQs are going up. Kids are -- kids know more, they can do more than they could 10 years ago or 15 years ago or 20 years ago and so you have to re-norm the test, and, you know, it's important to do that but then you're in a clinical trial, you really get into some hot water when you're moving from one to another. So there are some methods of taking that normative data in older children

and sort of correcting for the version -- some version control as a covariant.

DR. WAISBREN: And the last thing, too, there's a difference between a clinical trial and a longitudinal study.

DR. SHAPIRO: Yes, right.

DR. WAISBREN: And I think we should keep that in mind too.

DR. MULBERG: Yes sir.

DR. TSENG: Thank you. This is Brian Tseng from Boston. I'm a pediatric neurologist with Novartis. Professor Shapiro your slide that described the neurocognitive tools that you would recommend that they be short. Can you elaborate on that, particularly for young children? And I'd love to hear Mrs. Hogan's perspective when she mentioned the burden of participating in trials within neurocognitive assessments were, I assume, quite lengthy.

DR. SHAPIRO: Right, right.

DR. MULBERG: And if anyone from the panel would weigh in.

DR. SHAPIRO: Yes, we -- one of things that is most effective is making sure that you don't have fatigue affects. I mean fatigue is something that really occurs in very, very young children. So you need, for example, we use the Bayley and that

is the only test that we used in directly accessing the children in the Sanfilippo study and a Bayley can be done in a short period of time if you have skilled testers who know the test, know how to get that test done. I think Kate Delaney is going to talk about that this afternoon about how you do that. But, you know, in a natural history study, you know, you're not going to give two measures of cognition and you're not going to give a lot of other things that might be interfered with by fatigue.

DR. HOGAN: And to your point, I think this is where I talked about -- a little bit about being flexible at first and then being consistent because ideally when you have just using our case as an example neurocognitive testing with the DOS II and then kind of fine motor and gross motor skills. You know ideally even just the neurocognitive assessment would be broken up into maybe two different sessions because even sitting for that, you're spending a lot of time convincing them to stay still and convincing them to sit and convincing them to comply and not be distracted and there test fatigue by the end. So I think that's one thing that at the front end considering whether you can break that up, and I know there's certain requirements for that but then also within the test itself there has to be the need for flexibility based on the patient population. For example, one of the tasks relates to blocks in that test. And a

lot of these diseases they perseverate on certain things and so these boys continuously perseverate on stacking blocks and that was the first task. So if you left that as the first task, you were never going to see what they could do, so we moved it to the end and then you could actually see what they could do because you knew they were going to stack blocks. But I think being flexible there and figuring out within that population is going to be important.

DR. SHAPIRO: Yeah, I think that sometimes changing the order of the items will make things go more quickly and get the child engaged in the testing. And I think that, you know, breaking the testing up into segments and maybe even within the same day or the next day is another way of doing it. But making sure that you're not -- you know, one of the things that I think is very important in all these natural history studies is having a hypothesis, having an idea about what you want to measure. You know you're not going to measure everything and I think that was the problem that you had originally was you tried to measure everything and that's really not what a natural history study or a clinical trial might be. Now you might find that there, oh, we forgot to measure that and you might want to add that at some point when you see that that's a problem. But I think keeping it short is very, very important.

DR. BARON: I'm Ida Sue Baron. I'm a neuropsychologist. And I just wanted to make a small practical point about this decision-making about whether you switch tests or you continue with one. When a revision is made, it's not simply a matter of them getting new normative data. Items change and subtests are dropped and new items are brought in and the factor analysis might change. So it really is very difficult to think about it as only a change in normative data because that's not the problem. So if you really wanted to do an item count on the raw scores, your scores are about different topics sometimes. So, you know, and I've chosen to use, you know, the differential ability scale because it goes from two and half to 17 years 11 months. I just had decided I didn't want to have to deal with the Wexler and these scales that switch every four or five years.

DR. HOGAN: I was going to say she makes an important point because we noticed specifically in our tests, things that, in this day and age, kids have no idea about like typewriter or a lock or a mason jar. Things like that that my typical kids don't know what those are either.

DR. SHAPIRO: So I'd like to say something, and I think this -- Florian, this relates to your topic, and that is that the Wexler Test changed a lot over the years. And when we

originally were testing boys with ALD, we had the performance IQ on the Wexler -- I think it was the third addition, and it was extraordinarily sensitive to the visual problems that the children had and it was highly predictive. And then the test changed and the performance IQ no longer measured the same thing and it lost its sensitivity to the ALD kinds of difficulties that the kids had. But there was a subgroup of the PRI, which was more sensitive so, you know, these things are quite difficult to manage in a long-term trial.

DR. OPLER: Hello, Mark Opler. I'm at Columbia University, also, with a company called ProPhase; we do rater training and then point optimization test development, etcetera.

First, I want to applaud the FDA and the organizers. I think this is a phenomenal session and I'm privileged to be here today. But I want to draw attention to what I think is a little bit of a false dichotomy, so I'm going to throw out a statement, run away from it and see what everybody else does. And that has to do with, you know, the seeming dichotomy between what clinicians perceive, what caregivers perceive and what other observers perceive. You know, back when I was, you know, a young man, you know five years ago, and I was training in psychometrics, you know, we had a professor who kept going back to the idea that if you really want to get at the underlying

construct, if you want to get the truth about what's going on, you need different observers to do that. One observer is never enough. So to that end, you know, fast forward some years later, we're developing a new tool for a Down's syndrome program and it's got to be clinician administered and it's got to be short but it has to be ecologically valid and has to take the parents point of view into account. These are all the requirements that are being heaped on us by the sponsor, by other people, by other stake holders, by consultants and what we came up with, I don't know if it fit the bill or not, but I thought it was, you know, a reasonable compromise and that was a tool that actually required the clinician to write down on the form at the outset of the trial what the parent wanted to see their child achieve during the course of the study. And they then had to rate the child's progress on that anchor over the course of the study. I honestly believe that if we're a little bit more creative, if we try to get out of what we think of as these hard and fast categories between clinician rated outcomes, parent reported outcomes and patient-reported outcomes that we can come up with tools that are better fit, that get at again the need to have sometimes multiple observers looking at the same subject, the same patient to try to ultimately get at the truth. So I'm curious to hear if folks on the panel think that

we can get past this dichotomy, how we approach this, what it looks like from a scientific and operational and regulatory perspective. Now I'm going to walk away.

[laughter]

MALE SPEAKER: Let me try and take the first stab at that. First of all, I agree completely that we need the different perspectives. And sometimes what we find is that we're talking about exactly the same thing and using different language. And sometimes we find that one -- the priorities of one type of investigator or observer are different than another and sometimes we find we just don't agree on what we're seeing. I think, again, in cerebral palsy, in particular, there has been a move in some quarters to go to what's called a goal attainment scale, where at the entry into a trail, the goal is identified, and then the primary outcome measure is was that goal attained. Well, if you really ask the parents of children with severe cerebral palsy what their goals are, what they really want is their child to walk independently and talk independently and be able to hold a job and have a healthy, full life. With most of the interventions being tested right now, that's not realistic.

And so what that means then is somebody, a physical therapist, a physician, a nurse, a coordinator has to recalibrate that and steer them towards an attainable goal, a

realistically attainable goal, and now you've modified what the goal is. And so, yes I think that's -- that is -- that's what we all want. Is we want, you know, -- what are the ultimate goals about, we want to stop these diseases, we want to reverse these diseases, we want to cure these diseases but there are interim goals that are attainable, that are realistic, that are meaningful, and I think that, again, once you start to get to that…

The other area where, I see this all the time, is there are many children who have movement disorders that causes them to contort their face and the impression of untrained observers is that child is in pain and maybe in severe pain but if you actually ask the child -- no, there's no pain at all, they just can't control their facial movements.

Now, a lot of children you can't ask, so do you assume they're in pain or do you assume they have dystonia? I would say that we can't validly assume either but again it gets to that -- back to that area where we may observe the same thing, we may interpret it differently, we may give it different words or use different language to describe it and so it is a bit of a conundrum.

DR. OPLER: Just to react to that, I think you're absolutely right. I think you make a good point about the scale

on which we measure, right. So we also need to think about what we can measure within the context of what's often a study that's shorter than we would like constrained by resources, timing and the realities that we live in. So I think that's a good point.

DR. WAISBREN: You're talking about psychotherapy, also.

[laughter]

DR. WAISBREN: That's what a lot of this sounds like to me. This is how you deal with psychotherapy sessions and to translate that into a clinical trial, which they've tried to do for years, psychotherapy research, is very difficult.

DR. OPLER: I'm form New York; we don't know anything about psychotherapy there.

[laughter]

DR. HOGAN: And I would say as a parent I love the idea of that, you know, what is the parent's goal. I think the difficulty and challenge could be when you go into an interventional trial, even a Phase I trial, you may not know the potential of the therapy. So even setting that goal you might set it to low, you might set it way to high, you might set it the wrong domain so I think for some of these inborn errors it's so new in effecting the cognitive outcomes that we may have no idea how to set that goal.

DR. OPLER: That's a good point.

DR. BARBIER: And I want to add to this that if we move towards more of a goal attainment approach, first of all I agree that the goal still needs to be somewhat realistic; but, also we have to bear in mind for interventional trials, just the properties of what the drug can do especially biodistribution if a drug, oh I don't know it can get into the brain but not in the peripheral nerves then the expectations of what this drug could be expected to do need to titrated based on what, you know, pharmacokinetics, biodistribution, tissue penetration, [unintelligible], washout and all of those things. So it cannot become a Chinese menu of all the wonderful things you would like to have happen with your child.

DR. OPLER: Thank you very much.

FEMALE SPEAKER: Hi, I'm Jennifer, and I think this would be a great follow-up to the previous inquiry when there's talk of the truth and the realities what from patient perspectives on their missions and goals and I'd like to offer a couple comments in response to Ms. Hogan. In Dr. Eichler's previous remarks and from the perspective of the patient who has decades of history living with inborn error, specifically PKU. I was diagnosed in 1973 by newborn screening. And I'm very glad to hear the buzz for a shout out for newborn screening from a

couple of the previous speakers. I wanted to reinforce Ms. Hogan's statements about what's important to families because I champion that as well. One of them was regarding slowing disease progression and she drew attention to the need for improvement and that long path to bringing these treatments or effective therapies to market that really rise up and meet the challenges, the patients, parents, caregivers are faced. You want improvement in that long pathway to accelerate that and bring these innovations to the market, get the treatments in the patient's hands and look at it -- I can understand within the FDA, a scientific framework that balances the safety and rises up and meets the needs for the particular population. We're talking inborn errors that have lots of challenges regarding the biomarkers and looking at clinical parameters and their predictability and accessing all the outcomes and how they improve survival, affect function, the activities of daily livings. I wanted to really put a shout out for PKU in the adult population. I think there's so much historical data that we have at this point in time, that it could really set a precedent for future policy objectives and regulatory approaches to may be shift that framework. It seems almost like this county has a sickness model. Dr. Eichler made reference to having too much -- what you do with too much and to advanced

stages, you know, how about a preventive model and what are the lessons that we can learn from PKU. PKU is a recent example in which substantial prior clinical outcome data with a prior treatment was important for the use of the accelerator approval pathway to approve a novel drug, sapropterin. This was approved based on the use of blood phenylalanine levels as the biomarker endpoint in treating PKU. Blood phenylalanine levels were considered predictive of a clinical outcome based on public studies of intellectual outcomes in PKU during dietary therapy. Okay, the prior treatment that they're referring to was the dietary treatment, which in my case, and for many requires provision of medical food.

My question to the agency was I think you really need to rethink your approach to the category medical foods, specifically PKU and inborn errors. You have this division in the agency where we're at a meeting hosted by CDER that has purview of inborn error products and what I do not see is clarifying criteria on what distinguishes an inborn error product from a medical food, which is purview of the CFSIN that is used to treat inborn errors, PKU. We don't have this unity admission in addressing these needs for a population that is dependent on medical foods for survival on a day to day basis that literally save my life, save my kids' lives. You know, how

do you think that plays out outside the FDA world, when you have this division and fragmentation in authority and regulatory approach? It's a nightmare, and it translates into the policies outside the FDA world. And one of the biggest gaps that we are challenged with is, you know, bridging the gap and, you know, translating the discovery and to deliver it and getting access to those patients.

So I just wanted to draw your attention to that and perhaps, you know, they could follow up with me in one of the breaks, you know, or who I could follow up with to really, you know, -- I would just rethink that under a new paradigms and I would just try to ship more to. One of the lessons you can learn from PKU in setting new precedent, learning from, you know, newborn screening is one of the most successful preventative public health policies every. Why? It was developed because of PKU, and it saves thousands of lives. Thank you.

DR. MULBERG: Thank you so much for your passionate comments, and several of us would be glad to talk to you. Since we are a little bit off target, I wanted to make sure we get back on to the first session, if there are other speakers. Yes, sir.

MALE SPEAKER: Thank you very much. I just have a comment and then a question. The comment I had has to do with the power of the families in the investigative sites. In relationship to the Sanfilippo trial, and Elsa's to modest to be able to say that she's had a little debate about this, but when that study -- we wrote the protocols there, gosh 2009, it was envisioned as just a one year trial but as the families kind of banded together and to talk with the site and learn that they got a lot out of it, they actually came and said, "You know what, one year is not long enough. We know what you're trying to do, we feel empowered to do this longitudinal study right, we really need to go longer than a year." And, you know, as a pharmaceutical company, we were like, "Wow, yeah we agree and let's find a way to fund that."

And there's a second example of that; when we were trying to the Sanfilippo study outside of the U.S. in Europe, they have a different kind of guidelines. They have national ethics committees that said, "You know there's a measureable risk." Remember these kids had to be sedated for a period to be able to have cerebral spinal fluid taps and to have an MRI as they're kind of behaviorally disregulated and the government of the Netherlands said, "No, the risk benefit isn't enough to be able to perform this natural history study, we are not going to

let you do it." And the UK national ethics committee also said this and so we thought we're missing this opportunity to be able to do this and then the head of the MPS society said, Let me handle this.," and they went to the national ethics committee and said let me tell you about risk benefit. I understand what the study is; I've talked to the families here. I understand this is a preamble; the natural history study, it's an interventional trial. You really ought to let this happen and then of course it was approved in the UK after that. So you feel a little bit feeble as a pharmaceutical company but hey, you get the help where you can get it there. And it was really quite powerful. I'd like to say that had a great effect and then we enrolled in natural history study in the UK but the problem there was, and we've talked a little bit about it today, it's that -- it has to do with timing. Once there's an available treatment, it is absolutely impossible to enroll in a natural history study.

And so I think, you know, pharmaceutical companies that come to pre-IMD meeting and they talk about doing these things kind of concurrently, I mean it all sounds good theoretically, yes you can write those protocols but it's absolutely impossible to do that and Anne knows this, it happened at Shire several times and we heard some examples today

where a natural history study gets started, you collect a little bit of information, maybe cross sectional but once the therapy's available it just all goes away, which brings me to my question, maybe to Andrew Mulberg.

You know, when we wrote the natural history studies protocols in 2009, we really challenged ourselves to think about the studies as a potential for a no treatment or placebo arm in the study. And in some ways that makes a natural history study a much higher bar and even more difficult than thinking about a Phase I-II trail, which is really about safety but also collecting some information. The idea behind this is you want to collect with the same schedule of events, with the same rigor, with the same consistency across sites in a natural history study very early on, which is when you need to do it, before drug is available and thinking ahead in a kind of end of Phase II kind of discussions that you're having about will these measurements that we've done on the this schedule with this sort of central reading facility for MRIs etcetera, etcetera hold up as a possible, you know, no treatment arm or a placebo treatment arm which everybody wants. I think the families want it, you know, pharmaceutical companies want it, and I think the FDA would like to see that happen if they can be convinced that the

study then is going to be done in a very well controlled way, unbiased.

And so maybe I'll leave with a question for Andrew. I'm not asking for an answer to a specific Sanfilippo but just -- I'm sure you've had the discussions about thinking about the use of increasing more sophisticated natural history studies as a substitute for a placebo or no treatment arm in the pivotal trials. Just some thoughts on that.

DR. MULBERG: Well, by the nature of the question, I'll have to answer it with as general answer as you provided me with your introduction by saying I think we would be open to review.

You know, any of those novel ways of approaching difficult to recruit, you know, trials to answer efficacy questions. I think you've heard all the preliminary speakers from Dr. Buracchio to Dr. Goldsmith provide you some illustrations in the back drop. So clearly, what I would like to invigorate as a response to your question, is that I think we, as a community need to do more to answer precompetitive questions. To be able to avoid the kind of issues you're talking about. One thing we are not talking about in this meeting that is a critical topic for another meeting, is what do we do when we have two, three, four products for a disease that

has a 1 and 100,000 -- 1 and 1,000,000. This is a real issue and the community needs to talk about, so do we. Because we have our statutory requirements but then you have your operational constraints. So I would answer your question by saying the more we can do pre-competitively to address the common needs of the patients, the families and the clinicians to answer efficacy questions that we have to answer, that would be the best way to approaching it.

DR. SHAPIRO: I just want to make a comment about that one issue there and that has to do with these extraordinarily rare diseases and so many companies now out there competing for the same patients. And I think this -- it's very difficult and I think that this is something that in a sense is really problematic for the families when there are so many alternatives out there for the same disease and you, you know, you have a pool of 20, 30 patients, and that's it. Would you like to comment on that?

DR. HOGAN: Yeah, and I think this answers that question as well as the natural history and also addresses an item of significance which is the heterogeneity. And I think that's the importance possibly going forward of using historical controls, patient-to-patient historical controls, because you do have that heterogeneity. And if you have the ability to enroll

and you only have 20 patients you think that are going to enter this trial but you watch them for a year and then you see the intervention that's both inclusive of the community and what they want in terms of not having control arms but it's also very powerful to watch the drug in individual circumstances. So Dr. Mulberg may want to comment on the utility of that going forward.

DR. MULBERG: Yeah, again, that prospect always has potential pitfalls without a true understanding of what you underlined as the reason for doing it, which is the patient heterogeneity. Because you can make conclusions of successful patient 1 versus patient 2 and, you know, again I think this is not the topic for today's meeting but I will answer it by saying that, you know, we are exploring all potential innovative options with companies to pursue expediting drug development.

Dr. Parisi, you have the honor of the last question.

DR. PARISI: Goodness. Well, I wanted to present sort of a dilemma more than I guess a question but targeted towards Dr. Davidson and Dr. Eichler.

In talking about getting conditions added to the newborn screening panel, which is another area of interest for NIH, we've had this dilemma that the Secretary's Advisory Committee needs evidence that newborn screening, identification

or diagnosis in the newborn period actually improves outcomes. And the dilemma of course is that for many of these conditions that have a later onset of disease but of course, a progressive course, it's hard to accumulate the relevant evidence to show efficacy from a newborn screening approach.

So I'm just wondering if you had comments based on your experience, particularly with ALD, that might help inform this strategy to provide the evidence base for conditions to be added to the recommended panel.

DR. DAVIDSON: Florian, why don't you go first?

DR. EICHLER: So in the case of ALD, I mean, there's some very clear cut measures of replacing adrenal function that are live saving and I think recognizing those in addition to a good monitoring plan is crucial. I think your question is that, how do you accumulate the kind of data, you know, to where it's at. And I think, you know, the approach we've taken as a community is really to come together and bring all data together regardless whether they're accumulated by, you know, different centers across, you know, the world as well as, you know, patient experiences. And because the data is out there, it's often just not in that kind of, you know, the kind of form where it's readily translatable. But I would say in the case of ALD

it's -- the problem is more the lack of awareness about the clarity of the condition rather than the data itself.

DR. DAVIDSON: Yeah, I'd just add, from a pragmatic standpoint, for instance, in New York State, where newborn screening was recently instituted, they've identified a substantial number of affected patients substantial for a disease like ALD of course. And you know given the delayed onset of disease, there will of course be a lag before you're able to demonstrate the immediate clinical benefit and access to therapy that newborn screening would confer. But we do know today, and I referred to it and Florian referred to it, that fully half of boys today present with sporadic disease, they don't have a family history that enables intensive monitoring that would pick up progressive white matter disease and Gad enhancement that would then enable a transplant or trigger a transplant. They present with clinical myofestations, which are often markers of disease that's far to advanced to then enable therapy.

And so both in the case of preexisting therapy with allotransplant, I mean, newborn screening would allow for surveillance of effected boys to the point where they could get an early and timely transplant and hopefully with new therapeutics like gene therapy, it would enable a timely

implementation of those therapies as well. But it'll take time for this to play out.

DR. SHAPIRO: Don't you think that one of the big problems is this surveillance of the kids? You know, I used to see a lot of kinds with ALD and one of the problems was that, you know, you'd have a kid come in year after year and you're testing them and all of a sudden they disappear and then they show up three, four years later with disease to serious to test -- to treat. And so there needs to be some better monitoring of patients that -- it's expensive but it's lifesaving if you can do that. Just a comment.

DR. MINK: If I could just add on to that comment, too, because I think in New York State the experience with CRAB A was very expletive, that there was surveillance fatigue. And there was the dilemma, and this an issue I think for the funding agencies, that in New York State the CRAB A consortium declared it to be standard of care to do surveillance in a certain way because calling it research meant that New York Medicaid and the insurance companies would not pay for the cost of the surveillance and so it became standard of care in the absence of any evidence of what that surveillance should actually be. And in fact in CRAB A unless you have a 30KB deletion [unintelligible], we have no ability to predict whether you will

ever develop disease. And so I think that the surveillance is really the big piece of the problem and it's how do you pay for it and how do you structure it so that the families don't develop fatigue and so that you do pick up people when they are in that key early transition from presymptomatic to symptomatic.

DR. EICHLER: I'll add a final quick comment to this: For every boy that I diagnosis with symptoms who's too advanced, I find another two other family members who are presymptomatic and whose lives I -- who we can change. And I will say that the experience of a symptomatic boy who's declining is an incredible motivator for the family. And so while there -- it can be surveillance fatigue there is also this incredible motivation in patients and patient families that I think we have to harness. And I think this is happening and I think we really need to draw on patients also as educators.

DR. BURACCHIO: Okay. Thank you all very much for this very interesting discussion. With that, we will break for lunch, and we start up again at 1:00 with Session 3, where we will be discussing development of disease-specific scales.

[break]

SESSION III:

APPROACH TO ASSESSMENT OF COGNITION AND BEHAVIOR IN IEM

DR. BURACCHIO: All right, we're going to start the Session 3 right now. And our chairs for this session are Susan Waisbren, who you may remember from this morning's talk on urea cycle disorders. And we also have Peter Como, who is a neuropsychologist who works in the Center for Diagnostics and Radiological Health, or CDRH, here at the FDA. So, they're going to help moderate and keep things running for this first afternoon session.

And we're going to start off with our first speaker for the afternoon, which is Alison Skrinar -- am I saying that right? -- [laughs] okay -- who works for Ultragenyx. And she is going to talk to us about thinking about [unintelligible] selection, and how do you develop scales or modify existing scales?

WHEN DO WE NEED DISEASE-SPECIFIC SCALES?

HOW DO WE DEVELOP THEM?

DR. SKRINAR: Thank you. I'd like to thank the agency for giving me the opportunity to be here today to talk to you a little bit about the work that I've been doing for many years now. I want to introduce this session and talk a little bit about how you decide when it's appropriate to use a disease-specific scale and how you go about initiating the process of developing one.

As a disclosure, I am a full-time employee of Ultragenyx Pharmaceutical. We are a small but rapidly growing biotech company founded by my mentor, Dr. Emil Kakkis. And we are located north of San Francisco. I don't want to bore you with details about myself, but I am a developmental child psychologist by training. I am currently heading up a group called CORE, which is Clinical Outcomes Research and Evaluation, within Ultragenyx. And the purpose of my work is really to help to design studies. I worked for Genzyme for many years and then worked for Enobia and now at Ultragenyx. And my work has always pretty much been focused on the design of studies for rare disease, selection of endpoints, and, really, development of endpoints. I've done customized disease-specific measures for

muscle strength, motor function, cognitive function, PROs, ClinROs, so this has really become a passion of mine. And I have dedicated my career to really helping companies not be afraid of developing endpoints. I think it's in everybody's best interest because I think it really does help us better understand the diseases and ultimately better understand the impact of the treatments that we're trying to develop.

So a lot of what I'm going to tell you right now is pretty obvious, but it's hard to implement unless you have the courage to really, you know, barrel through some of the perceived obstacles. So, the first step in really determining whether or not you need to develop a disease-specific instrument is to understand your population. And that involves going to the source. Having been -- you know, studied in academia and then going to industry, we tend to rely on each other a lot to find out who's doing what, what's been used before. And we don't spend a lot of time going to the source, which is obviously the patient. But knowing your patient population is really the only way to determine if what's already out there will work for you, or if you need to develop something that's specific to the disease that you're interested in.

And by going to the source and knowing your patient population, I don't mean going out and just jumping into

quantitative research, but really going in and doing qualitative research. And a lot of that is just something that's just so basic. It's just clinical observation, talking to patients, talking to families to learn about their experience with the disease and what impact it has on their daily life. But for me, it's always about not saying a word and just watching. Especially with the -- with the assessment of neurocognitive function, it's so important to me to watch a child in their natural environment with their parent. Do they attend? Do they comply? You know, how impulsive are they?

So it's just -- do they seem to have, you know, the ability to implement, you know, just simple instructions? And so, that tells you a lot about whether or not a measure that's out there that seemingly looks great for what you're trying to study -- whether or not it'll work in a trial setting or in a real-world setting because it really depends on that clinical observation. It will tell you in your gut whether something is really going to work or not.

So clinical observation is key. Watching the patient with their caregiver is key, and interviewing them whenever possible. But the key to doing this and doing it well is to start early in the clinical development process. And it's been so amazing, over the years, just to see how everyone recognizes

the importance of doing this, and now to have the resources within the agency to initiate these discussions early in the development process, prior to the IND process because the one thing that I will tell you, whether you used an existing measure or a disease -- developed your own measure: it takes time. And that's really the thing -- everybody's trying to rush forward, and then we go forward, and we don't have what we -- what we know we could do better on, which is designing and choosing the right endpoints.

So the patients and the families themselves will tell you what the right constructs are. And so, once you do this qualitative research that I'm talking about, you'll know what areas you want to target for the assessment and evaluation of change. And you have to rely very heavily on this information. Trust your gut: what you saw in the clinic, what you observed, and what you heard from patients and their families. And I think we all know, having worked in this field, that orphan diseases are multi-systemic, and they're heterogeneous. So, what does that mean? It is -- it's just a fact. A heterogeneous clinical presentation is going to give you a heterogeneous clinical response. And so, what works for -- one child may not have a problem; another child has a very bad problem in that area.

And you're going to have to study multiple constructs in order to figure out what a drug is really doing. You don't want to just pick one that everybody has in common and then ignore other things that might actually respond to treatment. So, it's pretty much inevitable that you're going to be looking at multiple measures to capture the constructs that are of interest. So, it's not just the development of one, but you may have to look at multiple measures to make sure that you're capturing the nuances of the disease and any potential impact of treatment.

So how do you select these measures? Well, you can follow the conventional approach of using existing measures. You know what your constructs are, you've talked to your patients, they told you what was clinically relevant, and then you find that there's an instrument out there that assesses that. Big mistake because whatever's been developed and validated for another disease is not validated for your disease. The minimum clinically important difference that's been established for another disease is not the minimum clinically important difference for your disease.

And so, even if you pick an existing measure -- there's nothing wrong with that, but you have to be aware that additional validation work is still required in order for you to

understand how that instrument is going to behave with your particular patient population. And so, what that means is that you have to pilot-test those measures. You have to make sure that they are reliable and valid, that they're feasible and appropriate for use in your population.

So, using existing measures is not necessarily easier than developing a disease-specific scale. Obviously I'm -- I've done both. And I think -- I don't prefer one over the other; it really is a situation- and disease-specific decision that you have to make. But it has to be a very thoughtful decision based on the information that you gather in the clinic from the patients and their managing physicians and their families.

When you develop a disease-specific scale, you really need to strive to make sure that that instrument is clinically sensible and directly relevant to your population. And the one thing I will say about developing a disease-specific measure, now that I've told you that it's not really that much slower than choosing an existing measure because you still have to validate, is that you have this unique ability to find a way to capture a treatment response. And you don't necessarily have the ability to do that with existing measures. So, this -- you can customize it to capture exactly what it is that you're trying to look for. You can make sure that it is assessing

every one of those constructs that you're interested in studying. And you don't necessarily have to sacrifice one for another. But again, significant development time is required.

But that ability -- if your goal is to assess the impact of treatment, you do have a unique ability to do that, to look at the potential for responsiveness because the one thing I'll say about existing measures is, they're great at characterizing a disease population. They're great at looking at levels of disease severity among patients. But are they going to be sensitive to the impact of an intervention? And that's where you can really get hung up.

So, pilot-testing your measures -- again, go back to the source. You did your qualitative research. You heard from the patients, from their families. You observed with your own eyes what the problems are. You came back, selected some measures based on the constructs you identified. Now you have to go back to the source. You have to pilot-test. And pilot-testing never, ever, ever should take place in Phase II. It's too late. In fact, pilot-testing is not a single stage. It's an iterative process. So, you may have to go back several times.

I think, now that there has become more and more emphasis on the use of outcome measures, on trying to determine

ways to capture treatment response, there are a growing number of organizations out there that have services that can help you with the development validation process. How do you take the guidelines that we are fortunate to now have and implement those in such a way that you can gather the evidence that you need to develop an instrument where you can potentially use it as a primary or secondary endpoint for labeling claims?

Secure IRB approval -- we're always, you know, at least on the industry side, so afraid to approach patients. There's so many restrictions. I always write a protocol, always write a consent form, and just find a patient organization or key opinion leaders and tell them exactly what it is that I'm trying to do because we all want the same thing. We're all actively seeking treatments for these rare diseases. And the worst thing you can do is have a therapeutic intervention that works, but you don't have the ability to detect that change.

So, I go, I get a central IRB, I write a protocol, I write a consent form, and I get access to the patients, with cooperation, obviously, from their families and their physicians. And I test. I pilot-test. I pilot-test, I test muscles, I test motor function, I test cognitive function, I try out ClinROs, I try out PROs, and you know, I've had such -- if you just explain what it is that you're trying to do, I've had

such great cooperation from physicians and patients and families.

But all of this work -- and like I said, it is an iterative process -- has to take place prior to Phase II. When you collect Phase II data, that really should be a refinement process. You should already know what your constructs are. You should already know how you're going to measure them. And what I do is, I use that data to really look at, you know, preparing for phase III -- assessing, you know, how sensitive the instrument is to change, refining my scoring system, and building on that validation process. But that validation process should already have started.

So my key, critical message, if, like -- if no one remembers another word that I say, please don't wait because just because there's a measure out there doesn't mean it's feasible, doesn't mean it's appropriate, doesn't mean it'll be reliable and valid for your group. So you need to give yourself enough time to learn that and make adjustments.

I'm just going to give you a quick example of one of the things that I'm working on right now and how I've actually customized an existing measure, and now I'm testing how good I was, or not, in a phase II setting. But I've done a lot of pilot work. So this is Glut1 deficiency syndrome. It is a

glucose transporter deficiency. It -- the disease was first described by Darryl De Vivo at Columbia University in 1991. And it's caused by a genetic defect that leads to glucose transporter defects, so the glucose is not getting across the blood-brain barrier. And of course, glucose is the primary source of energy for the brain, so the result is that the brain is starved for energy. And this leads to seizures, developmental delay, and movement disorders. It's approximately 3,000 to 7,000 patients in the U.S. And because they aren't able to transport the glucose, the only fuel source that they have are ketone bodies. So, the standard of care is ketogenic diet with -- also some AEDs are used, as well, to control seizures.

So, just like every other rare disease, the phenotype is variable. And that does complicate the ability to assess the disease. It has a classic type that has seizures and developmental delay, and that's about 80 to 90 percent of the population. But even within that, those -- you know, 80 to 90 percent of the population is having seizures and developmental delay, but within that, two-thirds of the patients have more than one type of seizure. So, you're going to assess, you know, types of seizures differently. The cognitive impairment -- again, broad spectrum, and there's ataxia and dystonia and other

motor deficits. And then there's a -- you know, a minority of the patients that don't have the seizures but do have the cognitive and movement disorder. And they tend to be older.

And so what happens is that you have this variable spectrum in terms of the phenotypic expression of the disease. But then, within that, you're chasing it across the lifespan because it's changing. So, children, you know, mostly have the seizures. Then the developmental delay starts to occur, or it's detected. And then it's -- you see the motor component starting to evolve, and that's ataxia that ultimately leads to dystonia. So you're trying to do a study of, you know, infants, you know, young children, older children, adults, and everybody has a different expression. So, this is the -- goes back to that idea where I said you might have to study multiple constructs in order to capture the true response to treatment in a -- in a particular population, unless you want to study such a narrow subset. So we're always put in a position where -- do we study a very -- as clinically homogeneous a group as we can? Or do we study the broadest spectrum and look at multiple constructs, so that we can really capture what a drug is doing to the entire population?

So I set out to determine how to assess cognitive ability in Glut1 with a goal of better understanding the

cognitive abilities that existed in those patients. And obviously, you know, there's several challenges to this. As I've said, there's a wide spectrum of deficits. The delay presents in childhood, continues through adulthood. And of course, we have no gold standard for the assessment of cognition in these types of disorders or even within seizure disorders.

In addition to that, the kids have expressive and receptive language deficits. So it's very unclear, when you're trying to administer these measures, whether the child has actually heard you, if they understand what the instructions are. And so, you really need a test that's pretty easy to administer, that's pretty easy to understand, that can give you a broad range with which to measure a deficit, that's nonverbal, that's appropriate for pediatric and adult.

And every time you think you have, you know, a handle on this, you encounter another one. I was at a workshop yesterday for Rett syndrome. And I started thinking about how the work that I've done on Glut1 would translate to that. And the kids don't have fine motor function. So, all of the stuff that I'm working on right now that involves -- I thought I had taken just every element out of this. And they have to touch a screen, and I'm thinking, "Gosh, I have -- now I have to design a push button because they don't have finger control." So, I'm

thinking, “Gosh, it’s just -- it’s always something, another challenge in the development of these -- of these instruments.”

But the desired outcome is always going to be to identify, you know, a reliable and a valid measure to assess cognition in this -- in this population. But again, it has to be appropriate for pediatrics and adults. It has to measure a broad spectrum. It has to be nonverbal. It has to be easy. And also, within rare disease, you have to be able to implement it globally. You know, we’re often in a position where we have to do multinational trials. So, then you have language issues, training issues. So just actually finding the measure is hard enough, but then rolling it out into a clinical trial setting where you can standardize the assessments in the hopes of aggregating that and analyzing that data to tell you about your population and your intervention is even more complex.

So, what I did was, I was looking at the Cantab system. And you know, I’m just -- it’s just an example of something that can be used and customized in a clinical trial setting. But it’s a computer-based cognitive assessment system that consists of a battery of tests. So you can look at [unintelligible] and motor speed. You can look at working memory, visual memory, and you know, working memory, and -- with the ability to manipulate within the environment.

So there are a number of disease indications have used this system. Everybody tweaks it a different way. It started out for use in geriatric populations with Parkinson's and Alzheimer's. It has since been used in a number of pediatric indications: ADD, intractable epilepsy, and Down syndrome. It's obviously a nonverbal measure. It takes 30 to 45 minutes to administer. I've actually customized it down lower than that because I wasn't able to get the kids to comply for 30 minutes. And I've also had to break it up into, you know, component pieces so that I could make sure that I was getting as much of the child as I could. It has a standardized script and administration recommendations, and it has an e-learning module, which is amazing when you're trying to train and implement in global studies.

So I found that I was actually able to tailor this to work with the vendor that develops this, to tailor it to Glut1. So, I kind of ended up with some one-stop shopping, where I could look at attention and memory, executive function, decision-making, spatial reasoning, all in one place. But again, this has been an iterative process. I mean, just keep going back and tweaking it and tweaking it and tweaking it. And now, as I said, I'm testing it in phase II.

Age-matched normative data is available. I've had some success using it with children and adults. And it does accommodate a wide range of abilities. But again, customizing it so -- you know, constantly tailoring it down to get to that lower end of cognitive ability has been, you know, the biggest challenge.

So I -- there's a full battery of tests. The ones that I've selected were based on the information that I gathered from the patients, from clinical observation, from the families. It starts with a familiarization module; it's called the motor control test. So this is like a screening test. It's only two minutes long, max. But it basically tells you whether or not you can even try this with a child. And I've had a couple that, you know, just couldn't even get through this first basic module. But at least it told me not to continue to try to do that. I wasn't going to be able to get reliable data. But that's, you know, definitely been the exception and not the norm.

If they get through this motor control kind of screening test, if you will, I've been administering the reaction time test to look at attention and processing speed; the spatial span test to look at visual-spatial memory; the paired associates learning test, which looks at episodic memory;

and the spatial working memory test to look at executive function, and that has really proven to be too complex for this patient population. But it was worth a try because there were some, you know, higher-functioning adults. And I didn't want to see a ceiling effect. I knew that there were some deficits present, so I knew I had some area to work with.

The nice thing about the way that this test is structured is it starts with the easier tests, and they get harder. And then, within each test, it goes easy to hard. And I'll go through a -- you know, some specifics about each module. But the nice thing about it is that it will cut off. So, the child isn't set up for failure. And as soon as they miss a couple, and it's clear that they're not going to be able to get to the next level, you move on to the next test. So, they get re-engaged and re-stimulated, and it's not as frustrating.

So this just gives you an idea about --you know, when you train, you know, you're able to work with the child and do practice modules with them so that you're actually teaching them how to do the test. And that helps because when you don't have that language component and you have those expressive language disorders and some receptive problems, it's not -- you're not really sure if what you're saying and what you're showing is getting in there. So you're modeling it, and they're doing it

in response to what you're doing. So you can actually get a good feel as to whether or not the patient is really understanding the purpose of the test. It's all a lot of audio-visual feedback. And you know, I think it's stimulating for the child.

Again, I've shortened the test. But you know, for some of the kids, they can only get through one or two. Then we have to take a break, maybe the next day. But by and large, I've been trying to get it, you know, to a point where I could just sit down -- always looking at fatigue, but trying to make sure that I could actually get them to sit down and do the whole thing, if it's even, you know, possible. And I've shortened it in such a way that I feel like I've been able to capture the deficits but not fatigue them to the point where I have to stop and restart because I always worry about the impact that that has on the validity of the data.

[unintelligible] said, the -- it starts with the motor control task. This is administered at the screening visit just to familiarize the subject with the test system. It is a touch screen. And it assesses their cognitive and motor ability to actually perform the tests that are part of the battery. So, this does not result in any scoring, if you will. It can take, you know, two to three minutes, but the subject has to touch the

flashing X as it appears on the screen, and they have to do that 10 times. And you kind of just show them how to do it, and then they are actually modeling after you.

The reaction time test -- this has been a really good test with children that have behavioral disorders. And so, I've been able to look at vigilance and impulsivity with this test. So, they know if they look away, they will miss the opportunity to see the yellow box pop up. And so, it looks at their accuracy and their tendency toward a premature response. They have to hold a button down until the yellow dot appears, and then when it goes away, you know, they touch it. So, it can take up to seven minutes. I've been using the simple choice for the more delayed children and the five choice, which is, you know, a little bit harder. But it gives me, again, the ability to look at a broad range.

The spatial span test tests working memory, just like the Simon game, if you will. So you have to remember the sequence in which the squares light up with a color. And the number of boxes will increase, but you have to have, you know, success. So the sequence and the color varies. And you know, after three failed attempts, it will -- it will turn off, and you can go on to the next module. And to make it more complicated for the higher-functioning adults, you have to do it

in the reverse order. So, if one pops up on the top and the bottom, you pick the bottom and then the top. So, it will allow you to test the higher-functioning patients as well.

There's a paired associates learning that looks at episodic memory and new learning. And so, the -- each one of these boxes will have a symbol behind it, and then they'll go away, and one will appear in the center. And they have to remember where that particular symbol was. And the number of patterns would gradually increase until you have six failed attempts. I have, you know, tailored that down to three. So, all of the minimum failure requirements I've had to take town, and that's what I think has allowed me to get through the entire battery of tests with children.

So -- and then there's the spatial working memory test. And as I said, I can't even do this one [laughs], so it's really hard. But it does allow you to get to the upper range. And so, you know, I do have a demo model here, if anybody wants to take a look at it and actually play with it. It's kind of hard to explain how these things work. But I've found that it's stimulating, and the kids like it. And a lot of what it is based -- I didn't just -- it's not plug and play. You know, I've really put a lot of thought into what the specific deficits are in each one of these kids and then tried to apply that. And

if a patient doesn’t have a problem in a particular area, it definitely doesn’t count against them, but I am able to capture, you know, where the problem areas are.

So with -- this is kind of standard, but it’s really important, you know, and I think you’re going to talk this afternoon about standardization of the measures. But every one of the raters has to go through this e-learning training and certification. It’s done by the vendor; it can be done remotely. But it really encourages the rater to engage with the child first and establish a rapport with them in making them feel comfortable and setting the expectation that “This is going to become very hard. You won’t be able to do all of the items.” The test -- if the child is sensitive to noise, we turn down the volume on it. If they respond better to visual, then we turn up -- brighten up the colors.

And so, there’s a lot of ways to modify this, even within -- for an individual child, that will prevent, you know, some of that trauma that we were talking about, where these things are just so overwhelming. And again, it’s all clinical observation and gut. If it’s not going to happen, it’s not going to happen. And it’s not worth, you know, pushing too hard, especially when you know that, you know, you’re -- could be having -- you know, putting yourself in a position where that

child is just going to hate coming to the site, hate being tested, and really resist these types of measures. But all of those things, like the brightness and the sound, can be customized to fit the individual needs of the kids.

You have a testing script that you can follow. There's clear and simple instructions. You can model it. So, you can keep practicing until you're sure that the child, you know, understands what the test is. But all of the items that follow are novel, so you have -- there's not going to see, like, a training effect, if you will. And you know, the -- again, the most important thing is just to observe how they're doing it. You know, I think that I've learned a lot about what works. And there's -- there is a click button. There's also a space bar. I find the kids that have tracking difficulties do much better with the space bar. So, it just -- there are ways to build up the confidence of the rater that make it so that you're going to enhance the performance of the child. And some of the kids that have a little bit of, you know, stranger anxiety -- I've actually taught the mom, and the mom actually shows the child how to do the test.

And so, I feel like I'm getting good results. I -- hopefully we'll all be able to get back together, and I'll be able to show you how it all worked out. But I guess my main

point is not to be afraid to customize things. Get started early; play around with it. But really have the courage because I think it benefits everybody to figure out exactly what the deficits are, and it really drives the science. You know, we all want to know, with these interventions, if they're working and how they're working and how we can do better. We're just getting started with these rare diseases, and every treatment will build on the next one. And hopefully we'll get better and better outcomes for affected children and adults. But you know, it's really about capturing the disease, the nuances of the disease. And even in the absence of treatment, really being able to identify what those deficits are can help us implement treatment plans that, you know, can get us -- get these kids as functional as we can while we're waiting for interventions.

So, with that, I just summarize that I'm currently looking for Glut1 subjects, pediatric and adults, with a broad range. There really isn't any consistent measure of cognitive ability, especially when you're talking about trying to implement in multinational studies, where there are language issues. And the language deficits in the kids really complement -- complicate the administration of a lot of the existing batteries. But this is a nonverbal battery of tests. And the vendor that actually develops this is very amenable to

customizing it for different disease areas. But really, the onus is on you to understand the disease and then to reach out to them to figure out how to customize it. So you really have to have that person still on the front end getting as much information gathered about the particular disease.

So I'm right now in the middle of implementing this and hopefully will see good results. But like I said, I do have a demo system, and I encourage anybody that's interested in trying to do this -- you know, please reach out and talk to me because it's really exciting. And I think we can do a lot by taking, you know, the initiative to try to answer these questions early. Thank you.

[applause]

DR. COMO: So thank you, and we look forward to a discussion.

So our next block is about 45 minutes, in which we're going to learn from the experience of some of our distinguished speakers about lessons learned and how they went about developing their specific rating scales. We'll start off with inviting Dr. Mink back, from the University of Rochester, to discuss the development of the Unified Batten Disease Rating Scale.

STRATEGIES FOR DEVELOPING NEW SCALES OR MODIFYING EXISTING SCALES:

BATTEN CLINICAL RATING SCALES

DR. MINK: Good afternoon. Thank you. And that was a fantastic introduction to some of the things that I don't have time to talk about today. This is a disease-specific rating scale designed for Batten disease. And I'll talk about what that is and point out some of the aspects of what we've done and why.

Disclosures, none of which is relevant for anything I'm talking about today -- though thank you to the FDA for supporting a rare disease clinical trial -- I'm a pediatric neurologist. As I said this morning, I do a lot of different things, but I've been involved in research on Batten disease since about -- not 2012; since 2002. That's a typo. So, it's been about 12, 13 years now. Initially started with a conversation in the hallway with a biochemist who was doing biochemical and cellular research on one of the forms of Batten disease and had the goal of curing the disease and -- but needed -- if there was anything that was going to come to clinical trials in human beings, needed to have some kind of outcome measure.

Batten disease now refers to about 14 different diseases, we think, otherwise known as the neuronal ceroid lipofuscinoses, first described by Frederick Batten over 100 years ago now. Many forms have been described, initially defined by the age at onset, so classified by whether it starts in infancy or during the 2- to 5-year-old age range or juvenile period onset in the 4- to 10-year-old range or adult onset. At least 10 genes -- and again, there's debate about some others, but probably 14 different genes have been identified.

And we've learned some important things now that we have identified these genes is -- first of all, genotype does not predict the age at onset. And so there are different forms that may start during the infantile period or the juvenile period, depending on the mutation; that genotype doesn't reliably predict phenotype. There are some of these genes that have a broad spectrum of phenotypes and others, then -- the type I'm going to talk about most today, CLN3 or juvenile NCL, has tremendous phenotypic homogeneity, despite a diversity of mutations.

What they have in common is these are neurodegenerative diseases. They start in childhood, except those that don't. Those that start in adulthood are rare variants of this rare group of diseases. But together, all of

the NCLs together -- these represent the most common, the most prevalent neurodegenerative disease of childhood.

Almost every form has retinopathy. And so these children have varying degrees of visual impairment. But in CLN3 disease, they become blind -- they go from normal vision to being nearly completely blind in the matter of a year or two. So, it’s a very rapidly progressive loss of vision. And that, then, modifies everything that we have to do in assessing it because you have to have assessments that can be used in blind children.

Epilepsy is a prominent feature of almost every form of Batten disease; progressive loss of cognitive function, dementia, though it has somewhat different flavors in the different forms; movement disorder, though that also varies across forms; and there is this auto-fluorescent intracellular storage material that really defines the neuronal ceroid lipofuscinoses.

The differences you see listed there -- these are different diseases that have some important features in common. Here’s a shortlist listing nine of the first 10 that were described -- the convention is CLN, followed by the number -- initially in order of age at onset -- so infantile, late infantile, juvenile, adult -- and now subsequently in order of -

- in which it was named. So, these are identified proteins; some are not identified proteins. Some are soluble enzymes; some are transmembrane proteins. So, the therapeutic strategies for these different types may be very different, depending on the biochemistry. But I will argue that the assessment strategies and the strategies for both natural history research and developing clinical outcome measures actually probably have a lot in common across the different forms of Batten disease.

So I'm going to really focus on juvenile NCL, or CLN3 disease. It is the most prevalent, partly because it is one of the most slowly progressive. Usually starts between 4 and 7 years of age. Usually blindness or vision loss is the first feature. Death typically occurs in the third decade of life. And so, it's a gradually progressive but continuously progressive disorder. What that means is, if you're doing a clinical trial, you have to have some outcome measures that you don't have to wait 15 years to manifest. And that was the purpose of designing this rating scale.

So progressive blindness, cognitive decline, behavioral problems -- I know that makes Elsa Shapiro cringe. Specifically, they have anxiety. They have perseveration. They have other -- and they have aggressive behavior. So there are other, you know, specific things. But as a category, this

accompanies but is separable from the cognitive decline. They have seizures. They have a movement disorder, which in this disorder is primarily Parkinsonism.

And so we designed what we call the Unified Batten Disease Rating Scale with the goal of having something that was really going to unified, that would capture the main features of the different forms of Batten disease, and could be used for the different forms. But we designed it primarily for CLN3.

Initially, in -- the work was started by Fred Marshall about a year before I joined him. Initially what we did was, we identified items based on review of the literature for what are the key clinical features of JNCL. And what we found was two things. This -- most of what we thought about this disease was in the pre-gene era, and so there are things that were contaminating the knowledge about this particular disease. Second of all, the language that was used was awful, that the cognitive problems in these kids were described as either autism or schizophrenic. And that is -- neither one is the case.

So we developed some items, and then we went to the annual family meeting of the Batten Support and Research Association, where we had three neurologists each examine 23 children with juvenile Batten disease. Prior to that, each of us had maybe seen one patient in our life. This is a rare

disease. And what we learned was, some of the items we put on our rating scale were irrelevant, and that there are some things we completely missed -- so getting to the point that you have to go to the population if you're going to really develop good outcome measures.

So we did some formal item elimination and modification based on reliable -- reliability testing, both of the entire scale and of the individual items and until -- from 2002 to 2007 did continued assessment of the performance, reliability, and validity, guided by the data. And now we've had a fairly stable form since 2007.

How do you get the subjects? Well, we did the -- what we called the mountain and Mohammed strategy. We brought Mohammed to the mountain, and we brought the mountain to Mohammed. So we established a registry of known cases. We have gone every year to the annual Batten Disease Support and Research Association family meeting. And we established a clinical research center at the University of Rochester, where individuals come from all over the country to participate in our research or for clinical consultation, and then genotyped every subject so we knew that they actually had CLN3 disease. And so, we went to them, we brought them to us, and we tried to identify every known case in the -- in the U.S.

What is our scale? I'm not going to show you all the items, but we have demographic information, medical history, current and former medications. We have a physical assessment scale that is muscle tone and walking and trimmer and movement disorders and vision. We have a seizure assessment scale based on type, frequency, and severity of seizures. We have a behavioral assessment that is based on parent interview.

We have a capability assessment, and we do that in two ways. So the capability is things like play, school, self-care, other ADLs. But we ask the parents to give us an idea of what the child can do. And then we try to form our best opinion of what they would be able to do if they weren't blind. And there are some parents that say, "My child is blind, so I don't ask him to do anything, but he could do it." And there are other parents who say, "I don't care if he's blind; I'm going to have him do everything that's age-appropriate." And so we try to assess that separately. And it turns out, that works fairly well.

We identify the age at which the symptoms were first noted and try to get some idea of what the natural history is of the sequence of symptom onset. And then we also, for each of the core symptoms, have a clinical global impression score that

we've used to provide some additional validation testing of our scale.

Let me just show you a couple things we've learned from this -- is, first of all, we have -- I can show this in graphical form too, but basically, you can see that vision loss is the earliest symptom in both boys and girls. Behavior and cognitive problems follow a little bit -- behavior is a little bit more prominent early in the boys -- the cognitive problem -- and a little bit later in the girls. Seizures come on a year or two later. And then the motor problems come on a year or two after that. There are some differences between the boys and girls in age at onset, and I'll come back to that. But overall, the rate of progression is not different based on sex.

And this is just another way of showing it. This is the frequency distribution histogram of a Gaussian curve fit to the actual data. This is based on 93 subjects. The number of observations -- and it's a little hard to see the different colors, but the very high, narrow peak here is loss of vision. Some of the other symptoms have a little broader distribution. But we can develop a fairly good natural history profile in terms of when the different symptoms start.

Taking our physical assessment score, which is the one where we have -- this is the examiner, so this is based on

examination of the child rather than based on questionnaire. What we find is if -- there is a very nice monotonic relationship between physical impairments -- so high scores are bad, and zero is no signs or symptoms -- as a function of age. And we use age at testing as a surrogate for disease duration because they correlate so highly.

The blue dots are those who have -- are homozygous for the common deletion. And that's -- about 75 percent of alleles have the common deletion. And the red dots are those who have other genotypes; they're compound heterozygotes, or a couple of -- are homozygous for point mutations. And what you can see is that there's not much difference. And this is cross-sectional, so this is the score at the most recent evaluation. And this is longitudinal data shown here in B and C, where those individuals we've evaluated more than once have the different time points here with lines connecting them. You see, there's some up and down.

Makes the point, actually, that some children appear to be worse one time than they are another, and they appear to improve, and that depends entirely on how fatigued they are that day. It depends on how cooperative they are. It depends on all sorts of things; whether they had a seizure an hour before or not. But if you average those out over time, you can see that

there's a fairly consistent rate of progression. And there's no significant difference, except this one outlier here, who we still don't understand, between the two big classes of genotypes.

Our capability score -- 30 here is normal independent function, and zero is loss of all of those things -- also declines over age. And it's an independent measure, but what we see is there's a very nice negative correlation between our capability score and our physical score, providing some convergent validity that two completely independent types of measures that would be predicted to be related are indeed related, in terms of -- well, that the physical impairment is a -- is a major contributor to the functional capability.

We've recently looked at the function or performance of our scale using a telemedicine approach, where we compared an in-person assessment to one that was done by a trained rater observing an examination remotely. And this is a small number of individuals, but it's a very high inter-class correlation, so very high reliability, which means that even the thing that involves physical examination can be done remotely by televideo.

I'm not talking about the behavioral or cognitive stuff. Heather Adams, who is here and will be talking a little later, has worked to use other independent measures. And we've

found, again, some good agreement, convergent validity between what we assess using our rating scale and what is assessed using some previously validated scales.

One last point I want to make about this is that not only does this allow us to learn things about the natural history of the disorder, but it allows us to test some hypotheses. So we have now 12 years of data. We have some individuals we have seen every year for 12 years, and we have some individuals we've only seen once. But we've evaluated, now, over 105 individuals with this -- with CLN3 -- with the scale.

A few years ago, at the family meeting, a mother came to us. Her daughter had recently passed away from the disease. And she said, "You know, I noticed something. Of all the families I met when my daughter was first diagnosed, those who have boys -- the boys are doing better than the girls." And we found no data whatsoever to support a hypothesis that there was a difference in the rate of progression. But we did find that, on average, the vision loss occurred a year earlier in the boys. And then we looked at one of our measures, which was just a single component of our functional capability, which was when they went from "'needing assistance with ADLs" to "unable to do it without" -- so, unable to participate in ADLs. And so, we

used that as our -- as our marker, a dichotomous variable: When did they lose independent ADLs? And we compared the boys and the girls. And what we found was, on average, our 50th percentile here -- on average, girls lost independence in ADLs a year earlier than boys.

And so family input -- “I’ve noticed this” -- we went to our data. We could test the hypothesis. And yes, indeed, they seemed to have a more -- earlier loss of independence and a slightly more rapid course because the age at onset was a little bit later. And if we went to, now, the Batten Disease Support and Research Association registry, on average, girls and women with the disease die one year before the boys do. And so, there is a sex difference there that we never would have clued into, had it not been for this parent who said, “You know, there’s something I’ve noticed.” It also demonstrates that using this scale -- we can actually use it to test using prospectively collective data -- collected data to test some of these hypotheses.

Of course, what we want to use it for is as an outcome measure for clinical trials. But -- and we think that it has power for that, though we still have some work to do in terms of developing specific tests for specific milestones and specific clinically meaningful and patient- and family-meaningful

measures. But we have learned a lot about the quantitative phenotype and progression rate. We have a large data set now that we can use for testing these hypotheses and others. And it's been validated. It's reliable. We can use it with telemedicine, which means that in a rare disease, we can actually use this in clinical trials where we don't require the patient and their family, blind children in wheelchairs, to travel hundreds or thousands of miles to come to the clinical center where we have that expertise.

It does take a village. Some people say there are more people working on Batten disease than actually have it. Our team includes people who are not just at the University of Rochester but other places. Let me point out Heather Adams, who's going to talk a little bit later, is -- has been our neuropsychologist now for many years. Fred Marshall was really the instigator of this, having worked in the development of the Unified Huntington's Disease Rating Scale. And you can't do anything without statisticians or coordinators or students or patients. Thank you.

[applause]

DR. COMO: We're going to now welcome back Dr. Shapiro to talk about the development of the Sanfilippo Behavioral Rating Scale.

STRATEGIES FOR DEVELOPING NEW SCALES OR MODIFYING EXISTING SCALES:

SANFILIPPO BEHAVIOR RATING SCALE

DR. SHAPIRO: Hello again. I'm going to talk today about a newly developed Sanfilippo Behavior Rating Scale. And, let's see -- okay, these are same as before. So, Sanfilippo syndrome is a disease in which the behavioral abnormalities are perhaps the most severe of any of the lysosomal diseases -- maybe, maybe the most severe. I -- it's on that end. And we determined, when we were doing our natural history study, that we needed to find out what was going on with these kids. They had unusual symptoms. They were really different than other kids with other lysosomal diseases. And so, we decided that there was a need to develop a disease-specific scale that could be used to see if, in a clinical trial, you might see an improvement in behavior.

So how did -- I'm going to go through all the steps that we took to develop this scale. So, the first step was, we determined that there was a need for this scale. We read all the descriptive clinical literature. And everybody said they're aggressive, and they're difficult, and it's very nonspecific.

And we also talked to parents and asked them about the importance of the problem for that disease.

The next step was we observed the children with the disease, and we obtained parent and caretaker descriptions. And as you can see, the behavior was severe, qualitatively unusual, and interfered with parenting and with family activities.

So, we decided, then, that we needed to get somebody who was an expert in aggressive behavior and difficulties of childhood to come and observe these children. And so, we have a person at the University of Minnesota, Mike Potegal, who has just completed writing a big book on aggressive behavior in young children. He's a world expert in this. And so, I said, "Mike, will you come and -- to our -- one of our Minnesota meetings and come and observe these children in a natural setting and tell me what you think about these children?" And we also got some more parent reports of children. So -- and we obtained some videos of the kids. We observed the children with and without parents. And we got the parents to describe the behaviors that were most disruptive.

Well, Mike said to me, "Well, these children have Klüver-Bucy syndrome." And I said, "What?" I had -- you know, I had -- I knew what it was, but I somehow had not put two and two together. Klüver-Bucy syndrome has the following signs:

orality, repeated mouthing of objects, lack of social reciprocity, diminished fear. Diminished fear is the biggie. And it has been shown to be associated with amygdala dysfunction in many species, including nonhuman primates. It was first described with the ablation of the amygdala in rhesus monkeys.

So he said, “Well, let’s do a laboratory study of these children. On their first visit, when they come in for the natural history study, we’ll do a separate study, and we’ll look at them in a laboratory setting.” And there’s something called the Lab-TAB, which is the laboratory temperament assessment battery. And it has a whole series of different kinds of laboratory sort of procedures that you can do, that you can assess social interactions, fearlessness, startle reactions because startle disappears in Klüver-Bucy syndrome. And you can do staged encounters with people, objects, and loud noises. And we -- that’s called a risk room.

We also noted that these children had autistic-like behaviors. Many of the children had been previously diagnosed with autism. And so, we gave a standard measure of autism as well, called the Autism Diagnostic Observation Schedule. We got our autism experts to come in and do that.

This is what the risk room looks like. So, it’s quite the interesting experience. We have the children come into the

room, and we have a whole bunch of scary objects in there and a whole bunch of non-scary objects. You can see that there is a gorilla mask, a skeleton mask, a -- there's some nice toys over here, another skeleton mask over here. Oh, and over here is a stranger. Mike himself would come in, and he'd put a hoodie on and dark glasses. And he would sit there like this, with his hands like that. And what the children didn't know is, he also had his foot on a pedal. And when they approached the positive toys, these toys up here, there would be a very loud noise. And so, all of this was videotaped. The mother was in the room here, sitting here. And it was videotaped. And so, we could then have people rate the videotapes about the startle and all kinds of other interesting views of what these children did.

So one of the things that we found was that these -- well, going back to this -- that these children absolutely met all the criteria for Klüver-Bucy syndrome. They were fearless. They came into the room. They ran around. They touched the -- they were not afraid of the masks. They were -- they were noncompliant when the mother was asked to have them put toys in a basket; all kinds of things like that. We had them do all those things, and we compared them to MPS-1 kids of the same mental age, who would not even take one step into that room. They stood by their mothers the whole time. And when they were

finally brought over to the doll or to the truck, and they heard the startle, they all startled, every single one of them. And none of the kids over the age of 4, okay, in the Sanfilippo group startled. So it was only the kids under age four that showed some startle. And so, that was clearly a progressive thing.

So, from all of the data we collected from that and from the autism evaluation, we created items for a scale that was -- that were based on theory, the lab and the clinical results, observation, and parental report. And we scaled the items based on frequency, not severity -- how often did this occur, whatever the behavior was -- because judgments of severity are less reliable. Frequency is -- parents can do much better, and they're much more reliable about that. And it also accounts for normal children occasionally having these behaviors. The items had to be concrete. They had to be behavioral. They had to be understandable. And we also added, when did these behaviors start? And when did they disappear?

We then took the items on the scale. We organized them based on categories into domains such as: movement abnormalities; diminished fear; mood, anger, and aggression; social abnormalities; orality; masturbation -- that's one of the

behaviors seen in Klüver-Bucy syndrome; attention problems; hyperactivity; and so forth.

We then -- we had a 73-item scale. We gave the scale to the MPS Society. We sent them out to the parents of MPS 3 patients anonymously. We received 47 back. And on this scale, we also asked for volunteers for a telephone interview regarding the items. And so, based on the endorsements of the parents, we eliminated five items immediately were -- which were so infrequently endorsed that we didn't think that they were useful. And then Kate Delaney did 10 telephone open-ended interviews to explore the behaviors to determine if their descriptions concurred with the item endorsement.

So then we took all the items on the scale, and we calculated those items to measure the internal consistency. We developed four clusters by examining the associations of each domain with the overall predetermined cluster. It was movement, lack of fear, social-emotional abnormalities, and executive functions, orality and mood, anger and aggression. And they -- well, orality and mood, anger and aggression became standalone domains because they didn't cluster with anything else.

Then there's a statistical technique called Loess scores that examine the trajectory of mean scores across the age range in the study. And that's what these are. So what you can

see here is that some of the behaviors, like lack of fear, increases over time. Some of the items, like mood, anger, and aggression, seem to increase and then level off. So, what you get is an idea about what the trajectory is for each one of these items across time.

And then we validated the behavior rating scale with the ADOS, the Vineland, and amygdala volumes. And we looked at average change in the rating scale per 10-point difference in ADOS scores, Vineland score, or 10 percent reduction in amygdala volumes. And those are the associated p-values.

So what we found was that some of the scales correlated very nicely with the ADOS: movement, social-emotional dysfunction, lack of fear, not executive function, maybe a little bit on orality, and not mood, anger, and aggression. And so, we also found relationships with the Vineland, except in the movement area. And with amygdala volume, the most interesting thing was the lack of fear. And so, the lack of fear showed a significant correlation with manually traced amygdala volumes on the MRIs that all these patients had.

We then did a -- we were doing a natural history study with MPS-3B. So we did all of this original work with MPS3A, and then we moved on to 3B. And we found a very similar pattern in MPS-3B as we found in 3A. I -- yes. And we found that they

were slightly more fearful than the 3A patients. They didn't have the whole sort of intense Klüver-Bucy syndrome.

But the most interesting thing was that when we asked parents what the age of onset was of these various behavioral symptoms -- now, granted, this is retrospective data -- we found a differing age of onset with the 3B patients. And they first demonstrated the symptoms at a later age than the 3A patients, which is somewhat consistent with what we had seen clinically.

So what we haven't done, okay -- we haven't done test-retest reliability. And given the rapidity of their downhill course, it would have needed to be within a week because we found such major changes over a period of a month, for example, in many of these children. We need -- this needs to be done. This is a new scale. There are things that remain to be done. We also need to obtain a normative sample. What would that tell us? I'm not sure. But that's what people think we should do. We should see what these -- the frequency of these behaviors in normal children. I think they don't exist in normal children, but we'll see. It'll get done. We also have a cognitively impaired sample that has been collected in the U.K., and we're hoping to be able to soon obtain that data for comparison.

So what did we learn from doing this scale? It's very hard to find financial support for a control study of this. And

so, we didn't have the support to do that. Shire did support the initial development of the behavioral testing, and we had some money from our lysosomal disease network, RDCRN grant. So it was a very expensive study to do. It's very time-intensive. We had -- we had learned that we had to be available during the time that the parent was completing the measures. They had many questions and suggestions and wanted to discuss their children's behavior with us in great detail.

We did learn that the behavioral phenotypes in MPS-3 are unique. And we are now finding that there are other conditions with unique phenotypes. We've been consulting with a Gaucher disease type 3 neuropathic disease in Egypt, which seems to have a unique behavioral phenotype of severe aggression and conduct problems. And we're helping them develop a strategy for creating a behavioral measure.

So, as I said, this is a -- one of these very intensive kinds of development issues. Developing a disease-specific scale is very time-intensive and support-intensive. And it's not done yet. So, these are all the people who were helpful. I would draw your attention -- we have recently published it, and it is available online at the University of Minnesota. Thank you.

[applause]

DR. COMO: Our last speaker before we have the panel is Dr. Gerry Cox from Genzyme, who's going to talk about the use of computerized adaptive testing.

STRATEGIES FOR DEVELOPING NEW SCALES OR MODIFYING EXISTING SCALES:

USE OF COMPUTERIZED ADAPTIVE TESTING (CAT) WITH MPS-HAQ

DR. COX: Good afternoon. Like to thank the organizers for inviting me to speak to you today, really about two different approaches to get at the same thing, which is physical functioning of individuals with MPS-1. One approach is a standardized paper questionnaire, which we call the MPS HAQ. The other is a computer adaptive test, which uses a very different approach. And as I go through my talk, I'll try to highlight some of the salient differences between them and where I think the PDM CAT really shines.

I work at Genzyme. And I've been in clinical development at Genzyme for the last 15 years. I'm trained as a pediatrician and still active as a clinical geneticist at Boston Children's Hospital. I was one of Susan's students years ago -- I won't say 38 years ago, but [laughs] -- and I've really been interested in trying to do something for children that we would see in the genetics clinic for which there was really no treatment available for many years. And it seemed like the lysosomal storage disorders in the late '90s to early 2000s was

really a ripe area to try to do something transformative. And that's how I wound up at Genzyme.

And since I've been at Genzyme, I realize that a lot of the clinical research that's been done on drug development is really for common diseases. And it seemed like every time we studied a new disease, we've started at ground zero. So, I really hope that we can introduce some innovation into studying rare diseases and make use of technology, like I'm going to describe with the CAT, for this group of disorders.

So, you heard from Elsa earlier, mucopolysaccharidosis Type 1 is a prototypical lysosomal storage disorder. It's quite rare. It has autosomal recessive inheritance. It's caused by an enzyme deficiency, alpha-L-iduronidase, that leads to the accumulation of certain glycosaminoglycans, dermatan and heparan sulfate, throughout the body. These are extra-cellular matrix molecules that surround virtually every cell in the body. They're particularly prevalent in connective tissue. And so, when you think of the pathophysiology, we're talking bones, joints, eyes, heart valves as really being the most affected tissues.

There's a lot of heterogeneity between patients, not only across the entire spectrum of the disease, but also within subgroups of the disease. There tends to be significant

morbidity, early mortality. There are treatments available. Transplant has been around since the early ‘80s; that’s usually reserved for children that have the most severe form of MPS-1, Hurler syndrome, with cognitive impairment. And transplant has been shown to prolong survival and maintain cognitive function. More recently -- I say recently relative to transplant, 2003 -- laronidase was approved as an enzyme replacement therapy for the systemic manifestations of the disease. But the enzyme doesn’t cross the blood-brain barrier.

This gives you an idea of the spectrum that we deal with in our clinical trials, where we can’t really pick and choose enough homogeneous patients for an individual trial. We really wind up having quite an assortment of patients, which poses a lot of challenges in terms of looking at a consistent treatment effect.

In a nutshell, on the left is a patient with Hurler syndrome who has both severe systemic disease as well as severe CNS disease. This patient will have, in the untreated state, neurodegeneration after their first couple years of life, with early death during childhood. On the right, Scheie syndrome, originally thought to be a different MPS disorder, but later shown to be just a milder version of Hurler syndrome. These individuals tend to have well-preserved CNS function and onset

of systemic disease typically in late childhood to early adulthood -- joint contractures, heart valve disease, corneal clouding. And in the middle, Hurler-Scheie patients with intermediate disease. They generally have pretty well-preserved cognitive function, although there may be some learning problems, and variable amounts of systemic disease progression.

And thinking about what -- how we assess physical functioning in these individuals, you can see they have a lot of co-morbidities as a result of their disease: airway obstruction causing sleep apnea and fatigue; restrictive lung disease, both from the skeleton as well as organomegaly pushing up below, limiting endurance; corneal clouding affecting their vision; on the right, joint contractures -- these are children who are trying to extend their arms above their head. You can see the -- for the shoulder extension, there's about a 90-degree limitation. Similarly, in the lower extremities, unable to fully extend the knees and the hips. And then, there could also be CNS involvement -- not only cognitive impairment, but also spinal cord compression from infiltration of the dura; that can also affect lower function.

So, back in 2002, 2003, when we conducted our original clinical trial of laronidase, enzyme replacement therapy, we had combinations of both active as well as reporter outcomes. We

used a six-minute walk test as our primary endpoint and showed improved walking ability that children in the trial. Six-minute walk test had never really been used before in this group of disorders. Had been used in COPD, pulmonary hypertension. It really assesses what a patient is able to do.

What we found is that, while it worked well in a short-term trial, that after patients do this every three to six months for years on end, there is a real element of boredom that sets in. We have children that sit at the starting line, and they refuse to move for six minutes, and they get a score of zero. [laughs] And it's really assessing one function, which is walking ability. It doesn't tell you what's going on with their upper extremities.

We had hoped that the questionnaires -- the adult and child versions of the health assessment questionnaires that had been developed for arthritis -- we had hoped that the patient reports or the parental reports would confirm what we saw with the active functional tests. Again, these are tests that are not actually measuring what a patient does. It's measuring what a patient says they do. And sometimes, what they say they do may not be the same as what they actually do. And what we found is that there was really no difference in our placebo control

trial with the patient-reported outcomes over the same duration that we saw an improvement in six-minute walk test.

And we thought long and hard about, you know, what could have been a reason for that. These are subjective tests. There's a lot of noise. Typically you need bigger studies in order to really see significant differences between groups. This is a questionnaire that was developed for a different disease; maybe it doesn't apply that well. And maybe the items just aren't sensitive enough. We did find, over time, that we saw some movement in this outcome measure with clinically meaningful changes. But it took a long time to show up.

So we were looking, at that time, for an alternative that might be more sensitive and responsive to treatment. And it was around that time that Allie was still at Genzyme. And so, we had thought about putting together a disease-specific health assessment questionnaire that would basically beef up the HAQ and the CHAQ and make it more appropriate and sensitive to change for MPS patients.

And it was a few years later that we spoke with Christine Lavery at the U.K. MPS Society. And their group was actually very interested in outcome measures. And they were planning to evaluate some European measures like the EQ5D in their patient population. And they talked to us about including

some other measures like the MPS HAQ that they'd heard about. And we had recently heard about the PDM CAT from Steve Haley at BU. And so, they put together a cross-sectional telephone survey study, just looking at the characteristics and properties of these five -- four measures. For the sake of time, I'm just going to discuss the MPS HAQ and the PDM CAT for the rest of my talk.

So the MPS HAQ -- this is a paper-based questionnaire that was really derived from the HAQ and CHAQ but, as I mentioned, beefed up items for self-care and mobility. They assess a lot of common day functions, so I would say that they're very clinically relevant and meaningful. The MPS HAQ has a scale for each item that ranges from one to 10 -- sorry, zero to 10, where zero is "not difficult at all," 10 is "extremely difficult." And it was intended for use in children above age 5. It consists of 52 questions. It's the same questions presented to the patients each time. And that's the way the paper questionnaires work. And it takes about half an hour to complete.

The PDM CAT is a computerized adaptive test that includes the mobility and self-care functional domains of an instrument -- paper-based instrument called the pediatric evaluation of disability inventory. And the computerized

adaptive test has some really interesting, and I think innovative, qualities. One is that it's based on item response theory, which both the items that are being assessed, as well as the patient's functioning, are assessed along the same metric. And so, as patients answer questions, it gives you a -- basically allows you to home in on their functional status in a very methodical way.

The computerized adaptive test doesn't ask every single question. In fact, you wouldn't want to. This particular instrument has 279 questions in a data bank. But what the computer does -- it uses a smart algorithm. It asks sort of a general lead-in question; then, depending on how the patient answers it, it takes them either higher or lower in terms of functional status. And then, based on their response, it asks the next question based on all the previous responses. So it allows the questions to be very directed and, in a very short number of questions, home in on the functional status of the patient very quickly. In fact, this has been validated in other clinical settings, where patients either answer every single question on paper, and then they take the computerized version, where they're maybe just answering five to 10 questions per domain. And they've arrived at very similar responses.

So these were two different instruments looking at self-care and mobility. And now I'm going to just show you -- this is just a typical example in the MPS HAQ, the way the questions are formatted. It's just a linear scale from zero to 10 or "unable to do," asking simple questions about dressing or walking. The PDM CAT has pictures. And instead of a scale, they have basically four questions about each task: either "child can't do it" or "it's very hard, requiring help," "a little hard, requiring a little help," or "very easy." And then, same with mobility -- things like transfers, walking, getting out of a car. So, the questions are formatted the same way.

And this is an example -- if you were to be administering this test, you'd find that after the first question, the computer, based on your response, guesses what your score would be. And then it asks another question, number two, and depending on how you answer that, it adjusts the score a bit. Item three, it'll adjust again. And each time it's adjusting, you could see the standard error getting narrower and narrower. And in fact, the way this test work, you could actually set the stopping criteria to be once the standard error reaches a certain threshold. You're done because you're not going to be getting, really, any more accurate than that. And

down the bottom, you can see that, you know, after five questions, the score was essentially the same as if all 85 items had been answered. So that's a big load, I would say, off of patients, being able to only answer five questions instead of 85.

So the study that the U.K. did -- they sent out a request to all their MPS-1 members, and about two-thirds of them participated, 69. And the majority of these individuals were in the pediatric age range. About two-thirds of them had Hurler syndrome. Just over half the responders were the parents. And about 60 percent of the patients had received a transplant at some point in time, and about 40 percent were receiving enzyme replacement therapy.

And when we look at the scores for the MPS HAQ, there's a 20-point maximum for two domains, self-care and mobility. Each one has a range from zero to 10. And you can see the total score is right around the middle, 10 out of 20, and is pretty equally contributed by both the self-care and mobility. The means and medians are fairly similar, indicating that there's, you know, no huge number of outliers. And the range -- you can see the scores pretty much encompass the whole scale, either from three to 20 for the total, or zero to 10 for the individual domains.

Now, the PDM CAT also is based on a score of 200, a hundred per domain. In this case, the numbers go the opposite direction. So, a score of 90 is just a little bit below the midpoint of the range. The distribution is a little tighter, based on the standard deviation. But the means and medians are pretty similar. And again, the contribution of each domain -- the scores are fairly similar. On the right are the percentile ratings. So the PDM CAT has actually been normed. And so, you can see that for both self-care and mobility, patients are operating about a fifth to 10th percentile of what the normal range would be.

And really, this study was just to see how well these instruments would capture the diversity of patients. And you can see, there's a very tight correlation between the MPS HAQ and PDM CAT, which was reassuring to us. They're both supposed to be measuring self-care and mobility; just different ways of getting at the same question.

You can see that the MPS HAQ really does span -- the scores span from zero to 20. In fact, there may be a little bit of a floor effect on the right, with the worst score being the 20. In contrast, the PDM CAT -- it's a little bit tighter. The maximum score was just under 140. Minimum score was around 50, suggesting that there would be movement or room for additional

movement, either higher or lower scores. And when we looked at each of the domains individually, again, there was a nice distribution across the entire range. Again, I think, a little bit of a floor effect on the MPS HAQ, but both seemed to correlate pretty well.

So what I've shown you is that, using two different approaches to try to arrive at more sensitive disease-specific information, one a paper-based questionnaire with questions targeted for the disease, the other being really a huge computer database from which very select questions are drawn to try to establish a patient's functional status -- I think both approaches work. I think, with technology, having that large item data back, the smart questioning of the algorithm, really does make this an efficient way to conduct patient report outcomes. It's less burdensome. Many of these questionnaires are completed online -- two minutes, five minutes, you're done, instead of a half-hour, 45 minutes, one hour. [unintelligible] questions.

We're moving forward. We're still interested in the MPS HAQ. Julie Eisengart and her colleagues at University of Minnesota are doing some additional revisions to it. We'd like to norm it and validate it as well. But I also just wanted to bring up the topic of CATs. I don't think anyone else was

mentioning them today. And being in the technology age, I think this is something that we should all look into in the future, as a way to try to decrease the burden to patients as we take them through all these clinical trials. Thank you.

[applause]

PANEL DISCUSSION AND Q-and-A

DR. COMO: So, if we could ask our speakers to come up, we'll have a panel discussion. I think we're running about 15 minutes, roughly 15 minutes behind schedule. So -- okay, so we'll have about a 20-minute panel discussion and then take a break. Okay? Oh, right, yes. So, in addition to our speakers, we also have our colleague, Dr. Papadopoulos, who's from our study endpoints and labeling development, our SEALD team here at FDA, who will also join the discussion. And I'll ask Susan to lead the discussion to get it going, and then we'll ask -- open it up to the audience.

DR. WAISBREN: I don't want -- oops -- I don't want to take time from you all. But if people want to start asking questions, please do. But I have a question to get us started. And this goes to Allie. My question is that -- how do you show that a significant increase of a few milliseconds on a computerized test is clinically meaningful? I know that's been a question I've been asked many times.

DR. SKRINAR: I think, right now, it's -- all we can do is go based off of the age normative data. And there isn't a lot out there to put that into a context. So, I think that one of the things that'll be important over time is to look at how

that relates to activities of daily living that require a reaction time. So, it -- I think, right now, it'll probably be a requirement that it be complemented with some other type of more functional measure.

I mean, but you can think that reaction time -- you know, intuitively, if a child touches a hot stove, you know, obviously there are implications of having a quicker response time or, you know, being able to see a car and, you know, jump back up onto the curb versus being hit. But yeah, it's going to be tough to do that. So, that's kind of the downside of getting something that's super-sensitive in quantitative and then bringing it back up into a more clinical realm.

DR. WAISBREN: Thanks.

DR. SHAPIRO: Can I say something about that? On another test that we've used in older children, the TOVA, it has a reaction time measure. And one of the things that we've found is that we can look at -- we've looked at corpus callosum volumes and reaction time on that test. We actually also have Cantab measures. We haven't looked at that; that would be interesting. And we found interesting correlations between the TOVA reaction time and corpus callosum volumes. And we've been looking at that in the context of the idea that MPS disorders are dysmyelinating disorders, that the development of myelin is

not normal. And parents report this in terms of visual processing problems or slow processing in the classroom. So, we've done some correlations with that. So, I do think that those small reaction times do mean something from a clinical standpoint.

DR. MINK: And I would argue, reaction time is a biomarker every bit as much as corpus callosum volume is. It's a -- it's a biomarker.

FEMALE SPEAKER: Right.

DR. COMO: Questions from the audience?

DR. BARBIER: So we've heard and seen a couple of times that people show an excellent correlation of one test with another test or a cognitive assessment with a brain volume or a biomarker. And a lot of attention is focused on good correlations. But the question I'm asking is, maybe if these two parameters are not well correlated, is there a positive message in there as well? Is a potential explanation or interpretation that they measure two different things, and so a bad correlation might not be a bad thing? It might mean that you're on the track of two distinct aspects that need to be measured.

And as a corollary to that, is -- especially if you compare, let's say, one cognitive test with another cognitive

test, there's almost always implicit an assumption that one is the gold standard and the other is the -- is the proband, so to speak. How is that determined, other than based on just, you know, this is the test that has been used the longest in the most patient, and therefore, by default, it has become our gold standard? It's actually an open question to anybody who would like to bring that up.

DR. MINK: So, in our Batten disease scale, our physical assessment correlates very nicely with functional capability. But there is no correlation whatsoever between our physical assessment and our behavioral assessment, nor is there between physical and seizure. But there is between our behavioral assessment and the Achenbach Child Behavior Checklist, at least some aspects of it.

So, I think we need all of those things to decide, you know, for validity. You want to have some conversion -- evidence for convergence, where that's your a priori hypothesis. You want, also, divergence. And then, ultimately, if there's high correlation, then you don't do both measures unless there's a real strong reason to think they're measuring different things that are -- that are highly correlated. So, I agree with you. I think it's very important to have things that diverge, as long

as they are key parts of the disorder you're trying to measure, and as long as you have a good construct validity as well.

DR. SHAPIRO: We -- we've also found, in Hurler syndrome, that the physical -- the -- any measure of physical severity doesn't always correlate with the degree of cognitive impairment. Those are different dimensions. And, you know, we don't really expect them to correlate. We've looked at that to see if, you know, they would both be measures of disease severity. But they're really two points on a -- or two different dimensions on the phenotype.

DR. ADAMS: Hi, my name is Heather Adams. I'm a pediatric neuropsychologist in Rochester. And with Jon Mink and our team there, we study Batten disease. I think my questions are provoked by Dr. Skrinar's -- I apologize if I'm pronouncing your name wrong -- talk. But perhaps the whole panel can consider them.

First of all, with Batten disease, certainly, we see the issue of needing to be flexible yet standardized at the same time at administering evaluations of cognition and behavior. But within assessments and within and between subjects, I wonder about adjusting things like brightness or sound, and whether the dose being delivered of the stimulus changes, and therefore the psychometrics are impacted.

A second question I had that was provoked by your talk had to do with the issue of cognitive fatigue and cognitive stamina, and if perhaps there are ways for us to think of that as a variable in and of itself that's useful to evaluate. Thank you.

DR. SKRINAR: Oh, one thing I would like to clarify is that there were no changes made to sound or brightness in the trial. That was part of the pilot testing so that I could find a middle ground that would work for everyone because I didn't know how sensitive the children would be to light and sound. And so, I found that they were a little bit more overstimulated by some of the brightness, some of those pink colors, so we toned those down. But, you know, the point is to get everybody on the same page in the trial. So, it was part of the pilot testing. That's where I was trying to get rid of some of the noise and to see if some of this intense stimulation was actually keeping them from being able to go further into the test.

And as another point, we did hone down the items so that in the trial, there -- the kids are actually able to sit down and do the full battery of tests. So, I feel like I'm compromising a little bit. In an effort to standardize it, I

feel like I have, you know, made an effort to streamline it so that they can get through the whole thing.

DR. ADAMS: [inaudible] pilot testing is so essential [inaudible]. My comment about fatigue was not so much [inaudible]. That might be something [inaudible].

DR. SHAPIRO: Maybe Gerry [laughs] -- the revision of the HAQ is going to have a whole bunch of items on fatigue and endurance and -- the MPS HAQ. So, that's something we're well aware of, so -- I have a question for you, Alison. On the Cantab, were you using the normative data? Because if you're changing the number of items that you're using, that's going to impact the scores that you have because the normative data was gathered in a specific way, you know, with all the items and so forth and not discontinuing it at a certain point. And so, how are you handling that? Are you doing within-child changes or --

DR. SKRINAR: We don't have norms for this [unintelligible]. It was just to put it in context. So, it --

DR. SHAPIRO: Oh, you're not using the normative data. Okay.

DR. SKRINAR: Not for the trial data. But, you know, to characterize it, it was important. Then we started tweaking it. So, no, now that it has been modified, there -- it's -- the norms are not valid for that. But in terms of characterizing

the population, I think you have to do that first. I think you owe it to the community to characterize them and put them into the context of a normally developing child before you start tweaking and measuring and looking at the -- at particular intervention. So, now -- I mean, each patient will serve as their own control now that that instrument has been modified.

MALE SPEAKER: Hello. To trust the performance from some of these children, particularly with inborn errors, I've always been worried about. What about the cataracts? What about retinal dysfunction, which --we know that, I think, occurs in some proportion of the Glut1's, as well as so many of the inborn errors? So, how do you deal with that? I mean, do you do some screening tests? Can you exclude them? Could someone speak to that, please?

DR. SHAPIRO: I can speak to that in our MPS kids, who mostly all have corneal clouding. We do take the ophthalmological examinations to look at the severity of corneal clouding or any kind of visual abnormality, to make sure that kids are going to be able to see the test materials. We use a lot of visual materials in our older kids with MPS disorders. We have not found, actually, that unless there's a problem in contrast, that these children have any difficulty with the tests. We had some problems with the judgment of line

orientation, which was a test that we were using, because the -- there wasn't enough contrast. And -- but generally, the -- you know, their acuity is all right. It's just that the -- you need high-contrast materials for these children to be able to see the materials. So, we do check on that. And we haven't found it, but we always -- you always need to adapt your materials to any kind of visual or auditory handicap, which is also problem for these children.

MALE SPEAKER: Hello. Okay. So, first, one quick retrospective comment, I think, to my colleague from Shire. Correlations are great. I -- statistically and psychometrically, I prefer to see statistical agreement. Again, give me an ICC, give me a Cohen's kappa for preference. Correlations are descriptive, not analytic.

Now to my question. As we develop these disease-specific scales, what can we do to ensure their readiness for use in a multinational setting? I'll give you a very quick example. I'm not going to name the name of the pharmaceutical company that this related to, to protect the guilty. But, you know, one of the organizations we're working with -- we're sending stuff forward to be adapted for use in the U.K. And there, we're getting comments back from the ethics committees, the IRBs, saying, "There are too many Americanisms in this

scale." And we'd originally said, "Well, we need to spend time doing the correct kind of cultural validation." We were told, "We don't have the money. English is English. Send it on." And we were later, then, informed by the IRBs, "Too many Americanisms." What can we do now to make ourselves ready for the inevitable next step, and -- so that hardworking folks on my team don't get angry letters from the U.K. in the future? Thank you.

DR. COMO: Anybody want to tackle that? [laughs]

DR. SHAPIRO: [laughs] I guess I can -- I can -- I guess -- [laughs] I think Kate's going to talk a little bit about that later today. But, you know, when you do translations to other languages, you back-translate; you do all kinds of things. You need to do the same thing for use in other English-speaking countries. And so, the best thing to do would be to let the -- you know, to make sure that a group of people who are going to be using it are going to look at the scales and make sure that they're appropriate in that country. I don't know what you could do in retrospect, now that -- you know, you just need to take the scale, whatever it is, and make sure that it's appropriate for that country. Everything should be culture-specific.

DR. PAPADOPOULOS: That's a critical point, and one that we also often have to deal with patient-reported outcomes, that -- and there are standards for translation and cross-cultural validations that are used that has been published. So, you know, in addition to the forward and the backward translation, you know, the cultural piece is also critical. And so, doing some of the field work also within the country, piloting in the country, will be very valuable.

MALE SPEAKER: Thank you for that. And just to speak back to some of the comments, I think the linguistic validation is a very well-trod path. We all know how to do that. We've been through that a billion times. The bigger issue here is, sometimes there are culture-bound ideas embedded within the scales we create without us even knowing. We look -- we look at it; we can't see it, the same way fish can't see water. We need to start doing a better job at the outset, bringing in people who think, speak, and act differently, or as we call them, "foreigners." We need to start working with our international colleagues earlier on in the scale development process. End statement. Sorry.

DR. SKRINAR: I just wanted to say that we -- you absolutely have to do the cross-cultural validation. Forward and backward translation isn't acceptable anymore. It's very

expensive, and so, a lot of times, people think, “Well, that’s where you get the ‘English is English.’” So, it -- what it does is it puts the onus on you to, you know, get an idea of where your clinical trial sites are going to be so that you can start the process.

And the other thing is trying to find measures that actually have a lot of versions that have already had the cross-cultural validation performed. When you do a cross-cultural validation, you do have to secure permission from the developer of the instrument. And they have to be willing to do it. So, a lot of times, they have a version that’s been started or was abandoned. And they’ll collaborate with you. And you can get a discount from several agencies who are actually interested in promoting the use of global instruments. So, it’s expensive, and that’s why people avoid it. But it’s absolutely necessary, or your data will not be valid.

FEMALE SPEAKER: So, I have a -- I think a pretty quick question for Dr. Cox. On the PDM CAT, it has P for pediatric, yet it looks like it’s validated for a larger age range than just the pediatric group, or at least you administered it to patients with Scheie syndrome up to the age of 55. And I just wanted clarification of that because obviously it’d be really ideal to have a scale that you can use

across the age spectrum, particular for some of these Inborn Errors of Metabolism, where you have different ages of onset and disease progression and severity.

DR. COX: That's a good point. The original PD, I believe, was validated with norms up through age 7. Then the PDM CAT went up to age 14 with norms. But in fact, it has been used in older individuals, especially if they don't have completely normal functioning. And so, they're still going to be on the scale. There is another instrument, called a AMPAC, which is similar to the PDM CAT but for use in adults. And one of the efforts we had thought about at one point was actually linking those two scales because they look at very similar domains.

DR. SKRINAR: But the other thing I'll say to that point is that a lot of these scales are developed, and they max out at the age at which you should be able to complete a particular item. So, even though it goes up to a ceiling of 20 years, it applies to everybody because everybody has developed all of the gross motor abilities -- driving a car, doing all of those things. So, try not to get too hung up on it. It really depends on the items that are being asked. But once you're an adult, you're an adult. So, it's not really worth doing the -- all the extra norming to take you all the way out to age 65.

It’s just kind of like, these are all the things that evolve. And once you’ve stopped developing, you can really apply -- sometimes, depending upon the items, you can apply those norms for the 20-year-olds across the remainder of the lifespan.

MS. DELANEY: Dr. Skrinar, I have a -- oh, go ahead.

DR. COMO: I think we have -- we have time for maybe one or two more questions.

MS. DELANEY: Quick question about the Cantab. So we use it -- we’ve used it in the natural history studies at the University of Minnesota. I think it eliminates some of this -- the issue of cultural adaptation and language and translation, except for the instructions for the person doing the assessment. But I think, in picturing some of these kids who are with severe -- significant impairment, who either physically or cognitively are unable to do something like the Cantab -- I think it’s a great measure, and we’ve used it. We’ve used pieces of it, and we’ve selected those pieces knowing this is what they’re going to be able to do. Some things like spatial working memory are beyond their capacity. So, is there something in the works or in development for the more impaired populations that might be useful?

DR. SKRINAR: [inaudible] love to get a group of people together and to start thinking about this because I think

that -- I think that, you know, the development of these tools, you know, with the visual, spatial, and the audiovisual stimulation are really good, even for the very impaired. But again, like I said, I just went to a workshop for Rett syndrome and started thinking about the fact that I'm going to have to use another device for touch screen. So, it's -- the work needs to be done, especially as we're developing more products for more severely impaired children. So, be fun to get a group -- working group together to start working on that.

MS. DELANEY: And then a quick question for Dr. Cox about the six-minute walk test because I think a lot of studies are using that. And did you adapt -- I know the American Thoracic -- the developers of the tool are less restrictive than some of these others, Pearson and others, who have very strict guidelines about changing, and we have to worry about validity and standardization. Were instructions modified for these kids who were sitting on the [laughs] -- on the start line and wouldn't move? I'm thinking of some of the Hunter patients trying to do the six-minute walk test. You know, could we modify it so that we might get more data and keep it valid at the same time?

DR. COX: Yeah. We actually didn't modify it for the actual trial itself. So, we just try to adhere to the American Thoracic guidelines.

DR. SKRINAR: I've recently modified it to give them little bands, like, the little Lance Armstrong "Livestrong" bands. Every time they do another lap, they get another -- it's just a little trinket, but it was just a way to keep them motivated and keep them going. So, it's just -- and they have to put it back in the basket. So, they're approaching you, putting the bracelet in, taking out another one, and going back. It was just a way of keeping them motivated. But I mean, again, it invalidates it. But -- for -- with certain populations, you're just not going to -- I've done it. It's a boring test. So -- and my child wouldn't do it. So, I'm trying to be creative.

[laughter]

DR. COMO: I think I'll -- since we're running out of time, I'll ask the last maybe slightly provocative question because I think today we've heard two things. One is, the skill and empathy of the clinician or psychometrician administering a lot of these scales is paramount in these rare diseases. But now we're also hearing about computerized and computer-administered tests. And to date, other the nice voice of Siri,

I don’t know of any computer scales that have the clinical skills to administer these tests. So, I’m wondering if the panel would like to offer some comments about -- I know there’s pros and cons of both. But --

MALE SPEAKER: Siri [inaudible].

[laughter]

DR. COMO: -- I’d be interested in your thoughts.

DR. PAPADOPOULOS: I guess I can jump in. Just, you know, as a parent, my kids don’t have any problems with engaging in computer play and so on. And then as -- you know, from a clinical trial standpoint, if you have that sort of standardization, you then take out a little bit of that variability that we heard that can be introduced by the level of engagement of the investigator.

DR. SHAPIRO: I would say that computerized testing is probably okay for less impaired, older kids, maybe above age -- certainly above age 8, maybe even above age 6. But I think you -- you’re not going to be able to do it in children younger than that. I know the Cantab is normed down to five, but that’s a normal developing -- typically developing children. And I don’t think that most of the children under that age are going to be able to deal with a -- sort of the rigor of a computerized test.

DR. COX: So, you're talking about the performance-based testing.

DR. SHAPIRO: Yes. Right.

DR. COX: So, with something like the PDM CAT, that's actually validated down to birth if there's an observer who's answering it.

DR. SHAPIRO: No, that's -- no, that's fine. That's fine. Yeah, but that's an observer kind of [inaudible], but --

DR. COX: Yeah.

DR. MINK: There is a monkey Cantab for use in nonhuman primates that -- you have to train them, but no verbal cues are useful because monkeys don't understand what we say. And you have to give them a salient reward to keep them going. But I wonder, for -- and physical disability is going to have a different impact than cognitive. But I wonder, for the nonverbal kids, whether there is some way to use -- particularly if you're interested in things like, you know, spatial working memory and whatever, if --

DR. SHAPIRO: Well, so, some of the tasks that Adele Diamond has done -- you know, it's the same sort of thing, that the -- what is that, the one on the Cantab DMS task? And I think they probably could do that, but it's going to take a lot of training. And that's very time-intensive.

DR. MINK: [affirmative] And you have to standardize that.

DR. SHAPIRO: Yeah, and you have to standardize the training, and then that's going to take time and clinical trials. You need to have something that's fast and easy. And I don't think that's going to work.

DR. PARISI: This probably might not relate to these populations. But we've been using computer tests of different types of executive function for three-year-olds, and they really do enjoy it more than any of the other performance tests we give them. So, it really would depend on what, you know, level and the quality of their behavior. But we're finding them to be extremely useful. And some of them are being converted over, now, to tablet tests. So, there really is a whole new range of tests that have come out for preschoolers that a lot of people may not yet know about. And there --

DR. SHAPIRO: But the kids have to be able to point or to touch the screen --

DR. PARISI: I understand that. Right. I understand.

DR. SHAPIRO: -- and, you know, not miss, and that sort of thing.

DR. PARISI: Right.

DR. COMO: Well, I want to thank our panelists and speakers. We're going to take a 10-minute --

FEMALE SPEAKER: [inaudible]

DR. COMO: Okay. Please arrange to be back here in about 15 minutes, and we'll get our last session going. Thank you.

[break]

SESSION IV:

TOOLS TO STANDARDIZE ASSESSMENTS ACROSS MULTI-SITE TRIALS

DR. BURACCHIO: Okay, we're going to go ahead and get started with the fourth and final session for the afternoon. And in this session, we're going talk about tools or methods for standardization across cognitive testing in inborn errors. And we have, we've kept Peter Como as a chair [laughs]. We have a new chair for this session, which is Melissa Parisi, who is here with us from NICHD. And our first speaker is going to be Kate Delaney, and she is going to talk about some methods for standardizing neuropsychological testing.

METHODS TO IMPROVE STANDARDIZATION OF NEUROPSYCHOLOGICAL TESTING

MS. DELANEY: Hello. Thank you to the organizers for having me join the meeting and speak this afternoon. These are my disclosures: My background is in working at the University of Minnesota. I was -- I worked there for 17 years under the supervision and mentorship of Dr. Shapiro. And when I first started working at the university, I was working in a clinic, a clinical capacity, so working in our pediatric neuropsychology clinic doing evaluations of children with metabolic diseases and other things -- autism, attention deficit learning disabilities, and other neurobehavioral disorder. So I have this background in -- my expertise is in testing. That's what I've been doing for 17 years. I've spent hundreds of hours with families and their children doing assessments, so doing neurocognitive assessment. We have a Methods of Assessment paper from the natural history studies, the Sanfilippo studies that we did at Minnesota. We have -- we have a lysosomal disease network study of MPS-1, 2, and 6, of which I was involved. I had been involved for the past five years. We've done -- I've been -- done some scale development with Elsa and with others at the university. And in addition, I've been involved with training as a consultant for clinical trials with some of the sponsors

and involving MPS and other rare diseases. So that's my background.

We've heard today already all of the complications of doing these assessments in a standardized way. We've heard about the variety of tests that are available, the issues of doing these in a multinational setting, and the language translation/validation issues. We have -- for these trials, we have very few patients, so this is doing this in a standardized way. Finding people who are capable of doing these assessments is especially important because we have so few patients coming to the sites and participating in the trials. We can't afford to have lost data as a result of -- maybe, as a result of not having the facility or site or individuals doing assessments ready to see these children.

So we have -- one of the issues is we have a limited number of sites across the world who have the experience and expertise in doing assessments with these children. We have sites who have the patients, and we have the principal investigators who have the patients, but we don't necessarily have, we don't have as many people who know the diseases and know them well, and to work with these families and do the testing.

So I’m going to go back to that slide. The protocol and test selection considerations: We've heard a few times today the recommendation to start the -- designing the neurocognitive piece very early on in the trial and thinking about the sites who will be participating and talking to the principal investigators, if they're not the psychologists or the raters themselves, to talk with the principal investigators about getting their site ready and making sure they have the people at the site who are able to, who have experience with these diseases and doing this testing.

One of the things that Elsa and I have talked about is I think the disease -- experience with the disease is probably more important than -- experience with disease and testing is more important than experience with the tool itself. So all of the neuropsychologists in the room [laughs] -- good -- could add their opinion, I think. But I think that's probably more important. So having a clinician/rater who has experience with these children and their families and seen them either in a clinical capacity or in a -- or for a clinical trial for research purposes, I think that's more important than having someone who knows the Wexler tests inside and out, for example.

So, with these disease populations, we have a unique set of problems, and healthy kids who are developing normal --

normally are easier to test, and I'm not saying that a 2-year -- active 2-year-old is easy to test, but easier than some of these children where we see a lot of variability within disease. And an example I'll use is with the Sanfilippo Type A Natural History Study that we did at Minnesota, we had -- our experience was we had experience in the clinic with these children in doing development assessments, cognitive assessments with these children. So we thought we knew what we were -- what we would see in our research setting for the -- for the trial, for the natural history study.

And when we started, we had a handful of children come in for the first, for the baseline assessment, and I think after the third patient came in, I met with Elsa and told her that each one [laughs] was very different from the next. And that wasn't because one was 1 and one was 20 years old; we had a wide range of ages, also, but it was really the behavior where I saw a lot of variability. We had kids, we had participants, subjects, patients in the study who were -- we hear about the behavioral problems. They had behavior problems but not just hyperactivity and hyper locomotion, running around the room, throwing toys up and down on the table. But we also had children who were hypoactive. So, and that's not what we were expecting. So, we had these children come in who would sit at

the table, or sit on the floor next to me, or sit in their patents lap and, but they were hypoactive. And so it took a lot of effort to get them engaged in the testing. And we had one, there was one family who came in, and as we were walking into the -- and I'm sharing just a few stories only to emphasize the difficulty that's involved with doing these assessments and getting valid data.

We had a -- so we had a parent walk in to the testing room with me, and as we were walking in, she knew a little bit about what I would be doing. Her child was 8 years old at the time. I told her we'd be doing some things with blocks. Very impaired child with Sanfilippo Syndrome, so I knew we would be doing items from the Bailey [spelled phonetically], and I explained to her what we were, as we were walking down the hall gave her some insight, gave her some information about what we'd be doing. She became sort of teary, and she became apologetic and telling me as we were walking in, "He won't be able to do any of this. You know, I don't even know if we should do the testing because this is -- this will be impossible. He can't even -- you know, you'll see, but he can't -- he doesn't -- he doesn't pick something up off the table if I ask him to. You know, things like that."

So this is just an example. We sat down, and he -- I was wondering. I was concerned as well about his ability to participate in this assessment for this natural history study, for this trial. And he, after some time, he wasn't responding at all to what I was doing, and I was trying to think of everything that I, you know, things that he might be interested in, and this big bag of toys, the Bailey Kit that I had next to me on the floor. And I, I eventually, I was shuffling in the bag trying to find something. Because part of it, when you're in the situation with these parents, it's, for the person doing the assessment, when you see this, it's very stressful. The parents, the parents know what this means. It's extremely difficult for them to bring their child in for the, for the study, and then you're sitting there, and you're not getting anything. It's just what the mom said as we walked through the door: "We're not going to get anywhere with this."

So, I felt this, this need to do something. I didn't think we would get any usable data or information from the session, but I thought, "I need to get something. I need to do something." Because this -- it was becoming, you know, difficult, more and more difficult for the parent.

So anyway, I took out the mirror from the, this little mirror from the Bailey Kit, and this child lit up. He saw

himself in the mirror and he lit up, and he hit the mirror. It was up on the table. He hit the mirror. He didn't pick it up, but he hit it, and I held it up in front of his face, and he had this bright smile. So he was, in my mind I'm checking off the Bailey items because there are a couple of items with the mirror.

So my point is we had a lot of variability in behavior and skill level, and that's what we see with these children. So, in healthy kids you see this bag of toys: the Bailey or the Wexler tests, and it seems like these would be quite straightforward. But when we're, we're doing these assessments with kids with these diseases, it's very difficult.

We've talked a little bit about the, the cultural adaptations. Mark, Dr. Opler [spelled phonetically] was asking about the -- what do we do about these skills using them multi-nationally. And it is a problem. We have, I think there's been some progress made. Part of the issue I think is the test publishers themselves don't have clear guidelines on, on how that should be done. So the forward and backward translation might be acceptable to them, and then it becomes the official translation, and that's what's to be used across companies and trials. But I think there's been some progress, but it is an issue.

We have -- I think Mark's example was a mild one, probably, adapting English to -- U.S. English to U.K. English. But when we're talking about adapting U.S. English to Portuguese or -- you know, there are all sorts of other considerations. So, that's something to keep in mind when you're using these scales. Unfortunately, they were not developed to be used -- well, not developed for these diseases and not developed to be used. Mostly these are for the specific tests that I'm thinking of, they're really either normed on a U.S. population and were really designed in that way.

There are other measures. There are measures used all over the world that are similar to some of the ones that are more commonly used in the U.S., but part of the problem or part of the issue in planning these trials is we need to use something that everybody, that, that everybody can, that everybody can use. So, we can't have -- even if it's measuring the same thing. We, of course, can't have one test, the Griffiths or something, being used in the U.K. and the Bailey being used in the U.S.

So these are just some -- I've already talked about some of the things we've done to, or some of the considerations we have to keep in mind. We, specific considerations for these tests, and this comes with experience for the people doing these

assessments. Having, following standardization but also being able to exercise some flexibility. So, and this is test-specific and disease-specific, when you're seeing some of the -- you see these kids over and over, and you recognize some behaviors, recognize some of the more typical behaviors. But with the scales, you have to know it well enough so that you know when to be, when you can be flexible and when, if you change the, the order of testing, or you do something a little bit different from the instructions, that you're, that you're changing the standardization, and you're compromising the validity of the data. So, it's important that disease experience, but also being very familiar with the test itself and what you'll be using for the trial. So, this is important information for the sites.

And I would say that having, early on, thinking about a training program for the raters or the assessors who will be participating as investigators, as, as the evaluators for the study. So, and this is, this can be customized depending on the size of the trial, the number of patients you'll have, and the number of sites involved. There's a lot of effort and money, expense, that goes into training a rater, or there can be. And there should be a lot of effort put into that, but you can

imagine from the site's perspective, they may have one patient for the entire trial.

And so to spend, you know, dedicating 15 hours of their -- of their expert psychologist to this sort of training might be something they won't welcome. On the other hand, I interact with a lot of the raters who are doing the testing. We compare notes. We talk about difficulties we've had, especially on the Sanfilippo trials. We had, we had regular calls where we were discussing some of the issues that came up in testing, some of the things we used that worked and didn't work.

Some of the assessors that I've interacted with actually welcome a rigorous -- even though they have years of experience -- they welcome a more rigorous training opportunity. So, there is, some companies have utilized an online training, and others have -- there's been, the video training is utilized and has been has been utilized in trials for these rare diseases but others as well. And that's, the approach is the person doing the assessment submits a video for close and careful scrutiny by expert, expert examiners or psychologists who, who then return that information to the site, and there's some back and forth. That sort of training opportunity, there, some of the most valuable training opportunities that I've participated in have been at the investigators' meetings where you can sit

face-to-face with the assessors and talk about the issues that might come up and that have come up. But that's not always feasible. But that's another way in which I've been involved with training and found useful.

The follow up and support to sites that I'm just mentioning. With our natural history study of both Sanfilippo A and B, we had these, I mentioned that we had these regular calls. And that was really spearheaded by the sponsor. So, we had regular calls where we could talk about some of the issues that were coming up in testing these children. These are the obvious points of the importance of a good training program: getting valid data, recruitment and retaining patients. That ties together with a parent and caregiver satisfaction and participating in a study, so I had a, recently did an assessment of a MPS attenuated, a young woman with attenuated MPS-1, and she commented on -- she's very capable and commented on her experience in various settings and testing and how, and she was telling me, she, the participant, was telling me how important it was to know that people are well trained and that they're doing this; they can build rapport.

So, with the Natural history studies of Sanfilippo we had parents who were, seemed to be satisfied with their experience, and as Dr. Richard pointed out, they asked to come

back for another session, and that's not because they were so happy with what we were doing in the testing, in the testing situation. But we know that they were not, they didn't have a terrible experience. They were, they were motivated to extend that study for another reason, which was a scientific reason. We were happy to know that we must have done something right because they're willing to come back to Minnesota from Florida in January and participate in the study.

So, this is some repetition, but we want the assessors or raters to have a lot of experience, highly skilled. More than likely they'll have advanced degrees in psychology, neuropsychology, and if there are technicians at the site who are participating, they need to have experience and be trained and supervised by a psychologist at the site. So, you want to know that there's a lead supervising psychologist who, if they're not doing the testing, they're supervising the entire evaluation and on part of the study team.

So I've already mentioned some of these things, and others have as well today. Challenges are the medical status of the child; hearing and vision problems, which we see in a lot of these children; mild to severe or very severe cognitive impairment. If you miss your window of building rapport with the child or the parent, but especially the child, you might

lose that session, and with these trials we can't afford to even have, we can't afford to have a session lost because of rapport building. But sometimes that's difficult, and there will be times when we can't get data, for whatever reason. If this is a terrible day for the child, or the child has been sick, they traveled the day before, there will be times where we can't get that data.

These are things I've mentioned. We want direct training of the assessors, frequent communication with the assessors at the sites, so there's often the communication with the PIs at the sites. But I think also, especially when these are sometimes primary endpoints, making sure the people doing the assessments, the raters and clinicians at the sites, feel that, have some sort of support or point person to talk about protocol issues or test-specific testing situations. So, I've had calls from people who have said, "Well, this is what happened. Do you think this was -- why do you think this could have happened?" And they describe the child's behavior, and so having somebody like that, regardless of the size of the trial and how man sites you have, if you have more than one, I think having some communication, open communication with the people doing the assessments and the sponsor is a good idea.

So, these are the things, these are the things we can control for. So, I've talked about the problems and the difficulties around doing these assessments. We select the right tools that are appropriate for the population. We do training of the people doing the assessments. We do a screening process prior to the training to make sure they're qualified to do the testing. The data should be reviewed after it's collected at the site. And then some site-level considerations, just lessons learned, making sure that -- these seem obvious, but we have had situations where, "Yeah, well, he wears hearing aids but we didn't know he'd need them for this because we didn't know what we were doing here today."

So, you know, without that information the parent, the child might not want to, might not like to wear the hearing aid so they were left at the hotel because they didn't know what neuropsych meant, or, you know. So, just having all of these things; these are more for the sites' coordinators and site study managers to consider, but important things.

We've talked about this: giving results to the parents. With the natural history study, we were able to do that in other -- for these, for the clinical trials; the other trials, we're not giving results to the parents, and that's a whole hour's worth of discussion [laugh], I'm sure. I guess I

should say in some cases, results are not being given to the parents. Parents always know how their child is doing, but they also would like to see this. They would like to see it on paper, but they also have some understanding of why we're not giving them results on the day we're doing the assessment. The results might be given to them at the end of the trial. But it's important to, in building rapport and a relationship with the parent and explaining all of this and why the results are not given and the timing of assessment, this is something that can be easily standardized across sites all over the world with obvious things like travel, MRI, procedures in the OR. Again, these are obvious but things that have come up. So, don't do the neurocognitive assessment the day after the child has general anesthesia. Things like that.

This just talks about the issue when we can't get data. Sometimes it happens, and often it's a good experience or a good opportunity for reviewing the process of data collection. Also, when there's an independent data review or quality review in place for the study, that's a great way to get feedback about how the, how the assessors and the raters are doing in their testing. So, it's a, if you're doing that review you get some idea of how they're doing. You're not just getting the scores, but you're getting and idea of how they're doing or if there are

problems, and with that, you can go back to the site and to the raters to give them feedback on that assessment.

So I think I've talked about a lot of different topics, a lot of different things that make this difficult, and my summary is -- I didn't talk a lot about the tools themselves, but that's been touched on. But to have a plan in place for training, and even if these are qualified clinicians with years of experience, to make sure that they're qualified to do this sort of testing and that they have experience with the disease is also very important.

Thank you.

[applause]

DR. COMO: Our next speaker is Dr. Heather Adams from the University of Rochester, who's going to speak to us on the use of remote technology to expand the reach of clinical research.

USE OF REMOTE TECHNOLOGY TO EXPAND THE REACH OF CLINICAL RESEARCH

DR. ADAMS: So hello and thank you to the FDA for hosting this workshop. Thank you, also, to the excellent tech support today. So what I'm going to do is talk about remote technology to expand the reach of clinical research, and I'll preface all of this by saying that I view remote technology as a tool in our vast armamentarium. It's not the be-all and end-all to conducting clinical research, you know, as the new future thing we're all going to be doing for every assessment, but it's yet another tool we can add to our scope of assessment strategies.

So these are my disclosures, and I'm going to be talking a little bit just about some of the technical challenges that we've been dealing with -- maybe not technical challenges but travel challenges we've been dealing with, with our current phase II randomized trial. That study's ongoing so we really have nothing to report in terms of the trial itself. And then I'm going to be not presenting the data but just sort of discussing in general-terms data obtained from a telemedicine pilot study we did for a neuropsych assessment. Those data are

under review with a journal, so I wasn't really prepared to share them here yet, but I'll talk about them.

So I'm with the Batten Research Group in Rochester, New York. Jon Mink is the lead of that project, and I do neuropsychological evaluations and behavioral assessments and assessments of quality of life for that project. And so, my special interest in IM Disorder is an assessment of cognitive outcomes rises both out of my clinical and research hat as a neuropsychologist generally and then my research more particularly in Batten Disease. And I joined the team in 2003, in part because after the original folks in the group went to their first meeting in 2002 and realized there are a lot of cognitive and behavioral components and mood components to this disorder, they pulled me into help think about assessment of those concerns.

What I hope to do today, briefly, is to provide a rationale for a mode assessment in rare diseases; I'll focus mostly on cognitive assessment and behavioral assessments but will talk about other reasons and ways to think about it also. Give you a very brief overview of options for remote technologies for neurocognitive assessment. Talk about our experience in Rochester, and Dr. Mink has already mentioned some of this already, so that'll save me a bit of time. And then

perhaps just at the end provoke discussion for our panel later on some of the remaining challenges and considerations with remote technologies for both neuropsych assessment and beyond.

So, why should we do, or why should we think about remote assessment in rare diseases? Well, as many people today have already said, rare diseases are rare. We have small sample-sizes and individual cases are spread far apart from one another, so it's hard for investigators to get to patients and research participants, and it's even harder for our patients and participants to get to us.

There's tremendous cost in traveling to research sites or even for clinicians and researchers to travel to patients. We do home visits for some of our families who really are not able to travel or have a lot of kids, so it's hard to pack them all up together to travel. Access of course. Time, distance, expense, and we can't underestimate the physical burden upon these fragile individuals, especially as disease progresses over time. We ask a lot of people to take their child who is physically disabled and put them in a car, put them in a plane, and fly them across the country, or drive six hours to visit our center and participate in research. And they do it but it's a tremendous burden.

However, we know that there's limited local expertise in rare diseases, so even when you have people who are technically adept administering assessments, whether it's the Wexler or anything else, they may not be adept at understanding the disease and how to accommodate disease features in conducting a rigorous assessment.

And we think that -- and there are data already to show that when you can offer remote assessment as an option in some cases, that participant satisfaction will improve. And then perhaps following from that, you'll have even greater motivation and effort in conducting the assessment. And I would argue that we could potentially think about enhancing data quality as well. When you have remote assessment with individuals who are not just expert in the measures but expert in the disease and who can reduce the number of assessors, then you reduce, you reduce your sources of variability. And so, you can perhaps enhance your data quality, enhance the rigor of the data that you're getting.

So, disease experience really matters, and I was gratified to hear a number of people talk about that today. And there have been some really nice publications to address this. So, "Rater Training in Multi-Center Clinical Trials Issues and Recommendations," taken from that paper and I know it's small

text for you, so I'll just read it. "Rater should have enough clinical experience with patients who have the disorder being evaluated at all levels of severity, to recognize and judge the severity of each of the symptoms rated in the scale. Unfortunately some raters get little training before seeing patients in clinical trials and often learn on clinical-trial patients with clinical-trial data. Often it is their first exposure to patients with the disorder being studied."

And another paper by the same group, "Why Do Clinical Trials Fail? The Problem of Measurement Error in Clinical Trials: Time to Test New Paradigms." And these authors would argue that one of the things that potentially introduces unnecessary error variance into your assessment in trials is having multiple inexperienced raters. This leads to poor inter-rater reliability or interview quality and perhaps rater bias.

So, what is remote technology? Let's define it. Remote technology, remote assessment, I would say, is any assessment you can do where you're not in the room with the person. And that can be as low-tech as just sending a survey by mail and remotely asking either the participant directly or a proxy observer, such as a parent or a teacher or a remote clinician, to fill out a rating form and mail it back to you. And that has been done, in fact, in at least one study, to get

an estimate of cognition. And in that study, which was a study in Alzheimer's disease or actually it was a, it was a virtual clinical trial for minor cognitive impairment in Alzheimer's disease, and the investigators found that asking participants in the study and participants' partners or observer-rater partners -- so it could be a spouse or a family member -- that brief assessment of about 14 questions having to do with things like, "Recently, I find I have to write down things in order to remember them." Those self-ratings and family member ratings correlated pretty well with a direct neuropsychological assessment of the related domains.

Telephone assessment of cognition has been used for a number of years, both live assessment where you as the examiner call someone up on the phone and administer tasks. But also using interactive voice-response technology, which all of you are familiar with if you've ever called an airline to check on your flights, or you've called your credit card company to check on your balance. And those assessments can be done by something as simple as, "Press or say one" as your instruction.

Now we have with the Internet and improvements every year with the amount of data that can travel to you in the least amount of time, we have the ability to do really high-fidelity assessments without lags in sound or video or time. Doing a

live assessment via internet and webcam, I'll talk about some of the data we have in cognitive assessment in Batten Disease for that, and also Internet-enabled computerized tests, so the CANTAB is a computerized test, and there are other companies as well that have computerized batteries that have always been computerized batteries, and now these companies are looking at ways to push these assessments to the cloud so that you don't have to send a laptop or set up a laptop specially with those assessments, but providing that, you know, you meet tech requirements with the local equipment, you can have someone log onto a secure site and participate in the assessment in a cloud-based space, and then those data centrally are captured, again reducing error in data capture and analysis of those data on the back end.

And now, of course, as Dr. Barron [spelled phonetically] mentioned, we have smart phones and tablets, which were being used quite successfully in development of assessment of cognition, mood, motor function, and other domains.

So, there are a lot of options. Our experience in Rochester with Batten Disease is with some pilot studies, both for telemedicine assessment of physical function and now with cognitive function as well. Dr. Mink has already presented to you some background on juvenile Batten Disease, so I won't

really spend a lot of time on this except to say and underscore that Batten Disease is quite rare. We have about 130 kids genetically confirmed in our registry, and a smaller number of those kids participate on a regular basis. But we really try to make contact with as many kids as possible every year, but it's tough because we're all spread out across the country.

And this is an example: Our registry from 2002 to 2014 -- so it took us about 12 years to ascertain this number of genetically confirmed children. And the starts and circles that you see reflect Batten Disease Support and Research Association Center of Excellence, BDSRA-designated centers of excellence. Actually the site in Portland on the northwest coast has now been relocated to Texas, so we don't have any centers of excellence west of the Rockies. So you can imagine in a clinical trial where you have a rare disease and limited expertise amongst clinicians and researchers, how hard it would be for the families across the country to access that center of expertise. And Rochester is up there in the Great Lakes with six subjects in all of the state as of last year, and not very many of them were within a six-hour drive of us.

So, looking at this a little further, from 2005 to 2014, and I picked those years because in 2005 we began with some NIH funding to have the resource, the opportunity to fly

families in, or cover the cost of travel for families to participate in our research at our Batten Center. And so, over the course of these number of years, these nine or 10 years, we had 58 children who traveled to visit us from 51 families. Some families had multiple children affected. They made a total of 78 trips because some families came more than one time. We completed a total of 102 evaluations on these children, and the average roundtrip distance -- this would be for one trip to Rochester -- was 1,700 miles, with quite a large variability. So some people got in their car and drove a few hours; other people really were flying across country. There's the range.

So, let's look at that for our clinical trial. We're currently doing a phase II safety tolerability trial of mycophenolate, an immunosuppressant in juvenile Batten Disease. This is co-funded by the FDA and the Batten Disease Support Research Association. It's a 20-week trial, and in that trial the design looks like this: We have eight weeks in which children participate in one arm of the study. They're randomized either to placebo or to mycophenolate. The other name for it is Cellcept. And then they cross over, and after a washout period of four weeks they cross over to the other arm of the study. So, each child essentially is his or her own control. So, we don't have a separate parallel track of a

control group, but we were still able to build in a control arm in the study.

And so there are eight visits over 20 weeks. Four of those visits are in Rochester at the beginning and the end of arm. And then within each arm we have two local site visits. And so we have a wonderful study monitor named Sara Defendorf. She travels all over the country to find local sites. They could be local academic centers not too terribly far from where the family lives, or they could be private pediatric neurologist or pediatrician offices. And she essentially sets them up as "N of 1" clinical trial sites for us to conduct our safety and intra-monitoring four times over the course of the study. And we thought that that would really reduce burden for families because they wouldn't have to schlep to Rochester eight times in 20 weeks.

So how did that work out? So, for the visits to the local sites, there still was a lot of travel. And these numbers, instead of reflecting a single trip, this reflects the total collective number of miles that a family put under their feet in the course of completing this 20-week trial. So, by the time that all of our participants who are currently enrolled complete the study, they will have on average traveled about 310 miles to their local sites but with quite a wide range, from

eight miles -- so one family is lucky enough to have their doctor one mile away from them, and other families are traveling a total of 3,000 miles total just to get to their local site.

Now, how about the Rochester trips? Here we have over 18,000 miles under their feet as the maximum travel for one family. And so in the course of this trial, one family will have traveled three-quarters of the way around this globe, and this is with all of our efforts to have as little travel as possible.

So I think that remote assessment is something we should think about going forward because should we really ask this child, who may have problems with ambulation, who certainly is blind, who may require, who may require assistance, not just for ambulation but because they have cognitive difficulties and behavioral difficulties and also may be experiencing seizures from time to time. Should we ask this child to travel over 18,000 miles to participate in a 20-week safety intolerability trial?

So Dr. Mink talked about remote assessment in the physical exam. This just shows a little bit of how it looks. We have -- I'm not sure how to use the pointer, so I'll just kind of point with my finger. Dr. Augustine is in the lower right corner. She's the remote expert examiner. And then in

the main field you have somebody who, in this situation, is our nurse practitioner. But in our pilot study it was a medical student, and she was with the subject in the room and conducted the exam locally while Dr. Augustine remotely captured the assessments.

And Dr. Mink already showed this slide. Our correlation was extremely high, and when the components of the physical exam that require an examiner to be hands-on with the subject, we're removed from the analysis. The reliability did not really change. So, we found that by remotely watching we could still do a pretty good job of evaluating all components of the physical exam of the UPDRS.

We're starting to do this in cognition, too, and this top scenario reflects what you might perhaps see in a clinical, remote clinical evaluation. In this pilot study I was the expert clinician. I was located in another room down the hallway, in a hotel at the Annual Batten Family Meeting. And I was monitoring the assessment, co-recording the responses that were being captured by my graduate student who was the technician who was directly administering the neuropsychological tests with the subject down the hallway in another room, with a remote Internet linkup.

And I think this would be something you might see, potentially, in a clinical setting if you have a child who needs an evaluation for school or needs an evaluation in a local clinical psychologist's office, and you have someone who, again, is technically adept at giving the measures but doesn't have experience with the disease and just needs some coaching and hand-holding to get through it in the best way.

And the other scenario we looked at was a scenario that might more closely mirror a research, a clinical research multi-center trial scenario, where I, the expert clinician, am remotely administering the tests. I have experience with the disease, and I have training, and the clinical trial nuances of the, of the assessment. But we have a local technician who is able to just run interference: make sure the technology's working, can provide clarification if the sound is poor or if the child says something that the technician could hear but I could not.

And what we did is we piloted this a couple of years ago at a Batten meeting. We had three children with Batten disease and one child who was a healthy sibling control so we could see if there were any differences in the feasibility of doing this with kids with Batten Disease. This really was just a feasibility pilot project. And I don't have the data to show

you, but what I will tell you just verbally is that our between-rater agreement, so item by item on tests of verbal reasoning, attention -- these are all verbally administered tasks because these children are blind -- so digit span, vocabulary, information from the Wexler, wide-range assessment of memory and learning, the RAML recall, immediate recall and recognition, and verbal fluency that are between-rater agreement was generally between 85 to over 90 percent. And when we took out the kid who had, who was a healthy control, in some domains our agreement statistics went up. And they went up because the child who is healthy and socially engaged and not blind was distracted by the technology and by the whole fun of it. But our kids with Batten Disease who are used to doing this just kind of sat down in the chair, and they just got down to doing the work of participating in the study. I know that probably wouldn't happen with Sanfilippo Syndrome, Elsa.

So, to summarize, telemedicine administration of the UPDRS and the neurocognitive assessments were both reliable. Our raters obtained very similar answers whether they were in the room or they were down the hall. And so, we feel the physical presence of a trained rater may not be necessary in some instances. We felt that these administrations were feasible; they took no extra time than did the in-person rating,

and that's really important both for us and for the participants. The technology was not expensive, and it did not require extensive training in terms of the technology itself. And so the take-home message here is that it allowed for formal evaluation by expert clinical raters without requiring travel.

So here are some of those remaining considerations and opportunities for remote technologies for neurocognitive, neuropsychological, behavioral, mood assessment, and beyond. So, some considerations: The psychometrics really matter. It's great that we're doing this, but we need to make sure that the constructs and the things that we're evaluating are just as meaningful and useful to us from a measurement standpoint if we use remote methods. You have to think about the testing requirements of the test. You're not going to be able to administer block design remotely. So you have to think about different testing modalities for different types of domains and assessments.

And you also have to make sure your technology's adequate. The second year we tried to do additional pilot testing; we really had problems because the hotel we were in for the second year's assessment didn't have adequate bandwidth. And so, on the Friday when it was just, you know, us and the families at the meeting, we were doing great, and we had great

bandwidth, and I didn't have any lag issues. But on Saturday when there were lots more families at the hotel, and all the kids were on their, you know, on their cell phones and their video games and whatever and their online gaming, we had terrible, terrible connection, and we weren't able to really get any data. So, that's really important.

I would argue that it might be a little bit more extra work for the study team up front, simply because it's a new way or new model of doing things, but in the long run it's probably debatable whether or not it actually wastes time or not. And I think you'll find that, as an examiner, you're going to be a lot more efficient in being able to complete an assessment if someone can just hop right into the space virtually and participate with you rather than getting them in the room and getting them to sit down and, you know, getting them settled, and getting them comfortable and all of that piece. However, you do have to think about what type of environment you're doing this in. Do you ask somebody to travel to a local site to sit in a controlled setting, a nice, quiet room? Or do you do it in their living room with the dog barking and the doorbell ringing and all of that?

But there are lots of opportunities, too. I think there are tremendous cost and time savings for everybody. I

think you can boost participation effort and satisfaction all-around, and really, very, very important is to reduce burden for fragile subjects. And as I had mentioned before, you can perhaps enhance data quality as well. So, that's all neuropsychological, behavioral and so forth assessment. What about other things? Maybe there are roles for remote assessment when you're doing safety and interim evaluations. Doing your checklist of AEs, for example. Maybe you can look at outcomes assessment. Perhaps there are even opportunities to think about interventions. And I know that in behavior health studies now they're looking at ways to offer direct behavioral interventions remotely. Perhaps following approval, following the end of a trial, if there's going to be an observational extension study, there might be opportunities to continue following people to assess longer-term outcomes over time when it would be particularly burdensome to commit to having people come back in for direct assessments for years and years and years.

And recruitment, enrollment, and consent are also things to think about. We have an investigator in Rochester, Dr. Ray Dorsey, and he now is doing some studies where he obtains consent through a smart phone for minimal-risk studies, and it's working quite successfully. And it's I think probably improving engagement in research.

So, just some final words, and this comes from Amy Vierhile who you saw in that image. She's our nurse practitioner, and she's our major point of contact with families. When families get on the website, and they learn about Batten Disease, and they end up at our space, the phone number that they see is the number right to her desk. So this is what she says: "People need to get over the fear that it's not a good option," meaning telemedicine. "Some providers are very reluctant to engage in anything other than in-person, hands-on visits because it makes them uncomfortable. It may be less convenient for them as researchers but much more convenient for the family, and that's the bottom line." And also, it's always better for the patient if you can see them in their own environment where they're more relaxed, and I think for kids who have dementia, this is really important. If we ask a child to travel across the country to participate in a study, and the next morning they're jetlagged and they're out of their element, we're probably not going to get the best data from them that we possibly could.

So this is our team. The people whose names are highlighted in red were the ones whose work was really central to the information I presented today, but I know that none of

them would mind if I simply just thanked the families and the support groups who have enabled us to do this work. Thank you.

[applause]

DR. PARISI: Thank you. And our final speaker for this session will be Dr. Joanne Odenkirchen from the National Institute of Neurologic Disorders and Stroke at the NIH.

DEVELOPMENT OF COMMON DATA ELEMENTS

DR. ODENKIRCHEN: Thank you. So I'm going to talk to you about the NINDS Common Data Elements Project. Whoops. I pushed the wrong button. Okay, sorry about that. Okay.

I have no financial disclosures to report. I'm a clinical research project manager out of the Office of Clinical Research out of the Office of the Director. One of my main responsibilities is to oversee the NINDS Common Data Elements Project. And I participate in NIH-wide activities working with not-for-profit organizations in our national CD activities and data standards activities. I've got over 30 years' experience working with clinical trials and clinical research projects including data safety monitoring boards designing and overseeing clinical trials.

So this is the NINDS CD website, and really our purpose is to streamline your neuroscience clinical research projects. This is not something that we do but we want our clinical researchers to be involved in. So, why do we need CDs? And what is the overall impact of the NINDS CD project? It's to reduce time costs to develop data collection tools. It's to reduce study start-up time and the cost. It's to promote data collection in consistent format. It's to improve data quality

and to facilitate data sharing in comparison between studies and meta-analysis. And the reason why we started this project is about eight years ago we recognized that it took us about nine to 12 years to get our large, multi-site clinical trials up and going. And so, we started looking at about five of our stroke studies and two of our Parkinson studies, and we actually did some mapping of those studies, and we recognized that a lot of our projects had the same questions that were being asked but in a little different format. And so we mapped them, and we decided to bring a group of all of our data coordinating centers together and actually ask them, would basically doing some data common, doing some common data element project, would that be of interest to them? And they said, "Yes, it would be. And it would be good for our community as well."

So we have a huge collaborative effort in the CD development and implementation. We have expertise from hundreds of specialists around the world. This is not a U.S.-centric project. We've worked with over 900 people. They all volunteer their time and effort for this project. We work with other NIH institutes and federal agencies. We work with the National Library of Medicine. We harmonize with international data standards. We do work with CDISC to make sure that our, that our projects are put into the SDT in format. We work closely

with C-PATH as well. We actually have our NIH-funded studies. We asked them to put our CDEs into their, into their CRFs and actually in some of our studies we actually require them to use our CDEs. We also work with not-for-profit organizations and foundations, and actually some of them have actually used our CDEs in their registries and other projects as well.

So what are the objectives of the CDE project? It's to identify CDEs used in clinical research. It's things as simple as age, gender, race, et cetera. It's also to present data elements in a standard format available to all. We identify common definitions including permissible values, range checks, et cetera. We use standardized case report forms and other instruments. And we provide this information to researchers for data-use development. We don't actually develop the database, but we actually provide the information for people if they want to develop a database.

So this website -- or excuse me, this page actually provides all the information that you need to know. It provides you what a CDE is, which is the standardized question and potential answers. Allows for consistent collection of sharing of data, and it's actually the sematic value, the actual CDE name, in the clear definition. On the right hand side is an example, which is the CDE name, the birth head circumference

value. It provides you the definition. It provides you the data type, and it provides you the input restrictions. The middle is actually an example of what the case report from is, and the bottom is actually the CDE details. It's a lot of information on that page, and I apologize but it gives you a good example of what it all is on one page.

So, this is our terminology, our classification, and how we define our CDEs. At the bottom are our general-core CDEs. We have seven of those, which all of our disease areas are required to use. Those are things like age, gender, race, ethnicity, and education level. Doesn't matter what disease within neurology that you're going to use a CDE for. We require you to include that in all of our projects.

And then we have Disease Core CDEs. Those are things like if you're going to do a Parkinson's disease study. Doesn't matter if you're going to do a natural history study or a phase III clinical trial. We want you to include the UPDRS.

Supplemental highly recommends CDEs are those things that we would want you to consider to use for, let's say, a phase III clinical trial. There are certain things within a phase III clinical trial that we would want you to use, for instance, something for an imaging study. Supplemental CDEs are those things that may have not been validated in that disease

but may have been a validated instrument. And then exploratory CDEs are those things that have not been, that we kind of consider not quite right for prime time but may have been used in one, at one, in one study or by one laboratory, let's say for imaging or something like that. And we'd like -- people have recommended it, and we'd like other people to consider using it so it can be validated more.

So, how do we develop the recommendations for the clinical research CDEs? The processes to develop the research data sets are, we collect and review data report forms from clinical research projects and other outcome databases. We look at registries, clinical trials. We look at natural history studies and et cetera. We assess what can be shared between disorders, so we have over 10,000 CDEs already, and 500 instruments that have been recommended. So, we look at those. We don't reinvent the wheel, and so we look at what's already available out there.

We also identify working groups to assess what is needed and wanted in the core supplemental and exploratory sets. We develop data elements for clinical research use that may or may not include the outcome measures. We recognize that all of our disease areas may not be ready for outcome measures, so we don't require them if they're not ready yet.

We have a NINDS CDE team that develops a web-access training to get the group started. Now you'll say, "How does this work?" Well, let me tell you. Our CDEs are identified and developed and vetted by experts in the scientific community. We have a hands-off approach. We actually have usually a group of people who go out, we identify who do this. And we actually say, "Go do it." And I'll tell you a little bit more about that. The process is transparent and inclusive. We ask people from not only the U.S. but from around the world to do this. They are volunteers. We don't give them any money. We ask them to volunteer about an hour a month to do this as part of a working group.

The NINDS and the NINDS CDE team provides continuous support and guidance, but we really don't provide any scientific input into the recommendations that they make.

This is a slide that tells kind of the development steps in the process, in the timeframe. We have kind of a startup meeting, a conference call. Then we have, the working group is broken into subgroups of about eight to 10 people each. Then there's an internal review. And then we get feedback. Each of the subgroups get to look at everybody else's subgroup's recommendations, and then there's a public review comment. And then there are, in which it's posted for public review. And

then the subgroups get it back, and it's posted on our website for everyone to use.

Now, this process usually takes about 12 to 16 months, but we've had it, it has happened as short as six months. So, well, you say, "Okay, now it's done. Now what happens?" Well, version one is not the final version. We consider this dynamic, and it will evolve over time. The process is iterative, and we plan to annually review and update the CDEs. We also have an oversight committee that will, that will be formed to help maintain the disease-specific CDEs because we know that new instruments will be developed. We also understand that as new studies come about, there may be new data that needs to change certain CDEs that are already, that already have been developed.

So this is an example of the mitochondrial disease working group that just completed in February of this year. There were nine working groups. You can see what, what areas that they broken down into. And this is a charge of each disease-specific CDE working group. They were to complete a conflict-of-interest disclosure form because we believe it's important that if you're going to be part of a working group, you let other people know if you do have a conflict of interest. They met by teleconference. There were chairs that were

assigned to each working group, and they met as necessary, and then we provided administrative support.

These are the end products, and I'm not going to read them all. You can probably read them faster than I can actually read them out loud. But there were recommendations that we asked them to provide: templates, instructions, and a summary of recommendations.

I'm not sure if I slipped, okay, sorry, I thought I slipped.

We also, if they, actually in addition to the actual CDEs, if they wanted to provide recommendations on instruments, we actually provided a list of instruments that were similar to the ones that they wanted to, they might want to recommend. But we also collected additional instruments for them to review. And then we had them come up with a set of questions in which criterion, which they would look at those instruments. And I've got two examples of those to look at as well.

So this is from the mitochondrial disease working group. This is a list of questions that each group looked at for those instruments. And this is for the congenital muscular dystrophy working group that also completed in February of this year as well. So, a lot of the questions were the same about, you know, the population, what are the advantages,

disadvantages, the time requirements. Is there a cost associated with this? And should the instrument be recommended or not?

And then, finally, the additional steps for the CDEs. We do each ask, each disease working group basically to provide a publication, to write up kind of what their summary was overall. And then there is an oversight committee, again, to look at the additions or changes in the future, as the CDEs need to change over time as new clinical research data becomes available.

These are the products that are on our website once the working group is completed. This is what our website looks like for a disease area, and it's really hard to see, but I just wanted to give you a little bit of information. The light blue in the first paragraph, it says the start-up resource list. For each disease area it tells you, basically, if you want to start a clinical trial or clinical research project in any disease area that we have, that'll tell you what is needed when you start a clinical research project, what we require you to do, or what the working group recommended you to do.

Where it says, "Download Mitochondrial Disease CDE Recommendations," and it's got a little zip folder next to it, that gives you all the lists of recommendations from the working

group. Now, where it says NIH Resources, NIH believes that since we spent a lot of money developing the NIH toolbox, the patient-reported outcomes, and NINDS developed the quality of life, that Neuro-Qol, we want you to use those or tell us why you're not going to if you're going to do patient-reported outcomes or neurologic quality-of-life outcomes, so we put those up top. We're not necessarily going to make you, but we'd like to have that discussion with you if you're going to do at least a clinical trial funded by NINDS.

The next thing down is the overview of all the working group recommendations, and I'm going to show you what some of those look like. So here's an overview of the mitochondrial disease working group biomarkers recommendations. This is just one page of it, and it just gives a summary of what the overview is. Excuse me. This is also, this is guidelines document for the biomarkers working group. Again, I'm going to just go back real quick. This is the written document, the overview of the bio markers group, and now this is the guidelines. And basically they didn't have any core recommendations, anything that was required. But basically what they said is, "These are kind of the parameters. If you're going to do a biomarker study, these are the parameters for each type of biomarker, what you want to consider." It's kind of a menu. This was about 18

pages long, so if you ever want to do a biomarker study, there's so much information here for anybody who wants to do it.

So this is what the case report form -- this is just page one of 12. And again, no one would ever want to use 12 pages of biomarkers, but what you can do is within our website, we actually have a form builder. You can actually take this form, and you can actually develop your own form. You can reduce it to one page or only one question that you want. Whoops, this didn't show up right and I'm not sure why, but this is the CDE detailed report on or website. And actually the, this is the actual CDEs for each of the biomarkers, so you could actually take those and put these on our, you can actually take each row which is, like, the CDE 19548, and that will actually give you all the information to put, that you need to put in our database: all the permissible values, the range checks, anything you need to know about that CDE.

Now, this is a mega data. The permissible values are the mega data set that you need. We don't allow you to add to that, but you can reduce the number of permissible values that are in there, okay? So it's just something to consider.

So this is an overview of the mitochondrial disease working group cognitive behavioral and psychological group. These are the first two pages of it. I believe this is eight

pages long. The rest of the pages are just tables of all their recommendations, and you can see that the second page here has the domain cognitive. The sub domain talks about which, you know, if it's adoptive, emotional, behavioral, and then their classification there as well.

Now, each of these groups, they only looked at instruments. So this is the notice of copyright because we don't have CDEs for those because these are instruments. We're not allowed to put copyrighted instruments, the actual CDEs, on our website unless we get permission because it's against the law. And we get in trouble because we're the government, if we do that. So, basically we have this form that's called a Notice of Copyright, and we provide you all the information about it. Basically the classification, it's supplemental, a short description, the scoring of it, the reference, and actually where you can find that information. But we do contact the copyright holder, and we say, "We'd like to post it on our website and will you allow us to do that?" If they do, we will actually put the CDEs on our website. We haven't been very successful, but we do contact every single person, ever single company. But we're usually not successful. But this, right now, this is the way we go about doing that.

So I'm going to talk just for a minute about the NIH CDE activities because I am part of a NIH-wide working group that works on CDEs. We do have a web portal, and I'll show you that in a minute. We do work to reduce overlap redundancy and near misses. We are working with the Department of Health and Human Services on electric health records, and I'll just briefly talk about that. We do have a NIH CDE repository, and I'll show you a little, that in a minute as well.

So, as far as the electronic health records, we do match CDEs to standards in terminology required for meaningful use, especially [unintelligible] and RX norm, and if you don't know about that, that's going to be something that's coming down the road, especially with electronic health records. We have several NIH CDE initiatives now mapped to LOINC, the NINDS CDEs, both the core and the highly, and the supplemental highly recommended CDEs are mapped to LOINC. The NLM value sets is the authority center for quality measures. And there's a structured data-capture initiative and a few other things that are going on with electronic health records within NLM.

We do work closely with FDA, especially with regards to their therapeutic area standards in NLM, and the CDE working group is working with AHQR as well. We also work with CDISC and C-FAST on the therapeutic area standards, and actually, the

NINDS already has our Parkinson's CDE in the C-FAST CDISC format, and we're working with RTBI CDEs as well.

So this is what the NLM website looks like as far as the common data elements resource portal. It lists all the institutes and all the CDEs that are there. This is a few months old. We're in the process of updating it right now.

And as far as some of the funding announcements about CDEs, if people don't know this, but several of the institutes now have funding announcement that basically say you're encouraged, you're strongly encouraged, you're required to use CDEs. Within NINDS for a lot of our funding announcements it basically says you will use, and if it doesn't say in the FOA, it does say in your terms of award that you are required to use.

So, this is the NIH CDE repository, and it does have all of the NINDS CDEs, and several other institutes at least a good number of their, of their CDEs on the website, and it something that actually the department is looking at to put in some of their public health and safety data as well. And if you look at the bottom right, it has a create button. And actually we, NINDS, is working with them to develop a pilot -- we have a pilot project with them with the idea that we may be actually be using this as kind of a hybrid model in how we're going to develop our CDEs in the future. We've got some organizations

and some professional groups and some, and some disease organizations that we're working with right now that are actually creating CDEs in this way, and we think that this is maybe the way in which we'll develop our CDEs in the future.

So, real quickly we're going to -- I'll talk to you about how we're going to, how we compare our CDEs to FANX, which is another group out of NHGRI. They basically look at phenotypic and environmental exposures. They have 15 measures per domain. We have unlimited measures per domain. They've got 380 measures and they address 24 domains. They've also included sickle cell now, suicide, tobacco, and a few other institutes they've had some contracts, they've actually subcontracted with. They've got about 16,000 variables. We've got about 10,000 variables. They use the word protocol instead of kind of disease area or domain. They use the word data elements like we do. But we have a lot of thing similar, a few things different, but we do things a lot the same. They map all of their CDEs to NCI CADSR. We've got about 4,000 of our things mapped to CADSR. So, you know, I think in a lot of ways a lot of the institutes with NINDS are within NIH are doing things similarly. Some of them are a little bit different, but we are trying to kind of come together and map everything together.

So, what is the vision for NINDS in the future? All of the future NINDS-funded trials and large epi-studies will use the CDEs [unintelligible] CDE-compatible. As part of our FOIAs in the terms of awards, we think that all types of our clinical research projects can use part of the CDEs, including our observational. Clinical studies can be linked to trial data sets, and all human-subject grantees are asked to consider using our CDEs. Clinical research progress will be accelerated, new investigators can build on consensus data elements, and a start-up of multi-centered international clinical research efforts will be facilitated. And we are collaborating with the National Library of Medicine on a hybrid model.

So, to access our CDEs, here's the information on how to do that. You can provide feedback. There is a page on our website to do that as well. For information you can contact me at the email below. And just to acknowledge all the about 900 people who have worked on all of our working groups and our domain areas. We do have a steering committee that oversees all of our groups. We have an oversight committee for all of our 19 disease areas. And all the disease and disorder organizations and the professional organizations and the international collaborators I'd like to acknowledge as well. Thank you.

[applause]

DR. COMO: So if we could ask our speakers to come up. I'll just share as being part of the Huntington's Disease common data elements, those calls led to spirited discussions about what should be in the core, what should be supplemental, what should be exploratory, but I think at the end of the day, at least from my experience on the Huntington's Disease CDE, I thought the product turned out quite well so…

We would like to invite folks to come up and ask questions of our panel, and if not, we may have one or two.

PANEL DISCUSSION AND Q-and-A

DR. BARBIER: This is a question for Dr. Adams. So I'm really very intrigued by this idea that you do not have to go [unintelligible] this enormous logistical enterprise of having typically one parent with a very sick child go to the site, whereas the other parent, supported by extended family, friends, and neighbors, tries to take care of the rest of, you know, family life. We had considered this in for some of our trials to do videotaping, and then have that be scored separately. The feedback that we always got is that for some of these tests, especially the ones that take a longer time, the children don't necessarily sit still at the table, but they start roaming around, and so you cannot have a single camera that can follow everything. Also that there are so many nuances in terms of communication that cannot be picked up by the audio or can be picked up by the camera, or even specifically in the context of mucopolysaccharidosis, that the presence of a cameraman, you know, following with a camera, would just be an additional distractive element. So, I would like to hear your, your view on, for what types of tests does it work? For what types of tests is it perhaps more trouble than it's worth?

DR. ADAMS: So you've touched on something very interesting, which is the intersection of technology and phenotype and assessment domain. And all those things really need to come together as a good triad for remote assessment to make sense. They all need to work, and they all need to work well.

Some of the technological issues I think are being surmounted with every passing year as technology advances. You can do things like, you can mic the child so that you don't have to worry about whether or not you can hear the kid. Having a local technician in the room can help to assist with some of those camera-positioning issues, or clarification of something a child says which is hard to hear. In Batten Disease, in juvenile Batten Disease, over time kids lose the ability to clearly, so speech articulation is a concern and consideration. And so, I would view cognitive evaluations of children with Batten Disease as useful for early- and moderate-stage disease but certainly not for later-stage, more advanced disease. And as children are moving into the phase of the disease where their speech becomes poor, it's really helpful to have a technician in the room who's at least has a passing familiarity with the speech pattern and also has advanced training in not saying to you, as the remote examiner, anything more than what they heard

the child say, to really just provide exact verbatim repetition of what they heard. But you do get to a point where you can't do that anymore. So, that's one issue is the technology.

I think the other issue is the domain of assessment. I allude to that a little bit in my talk. You're not going to be able to give block design remotely in the same way. Perhaps you have a technician administering it locally, and you're remotely observing and coaching. But there may be some assessments that you remotely as the expert can administer over the video, and other assessments that you want a technician or locally trained person to administer directly, and you remotely are linking in to provide the coaching and support. So, I think that's another option, is that you don't necessarily have to give all tests from 1,000 miles away or further. You can have someone locally who has some training. But by having someone centralized who has some expertise, you may be able to facilitate a more rigorous assessment.

MR. RICHARD: I kind of have a follow-on question from Anne -- this is Charlie Richard -- for Heather and maybe Jonathan Mink or the Minnesota Group who's thought about this. This is a new world order of kind of gene therapy. So, you know, in the old days of -- and some replacement therapy where you're administering these complex biologicals every two weeks

or every week, patients have to come in anyway. You're looking at infusion-related reactions, and so thinking about tacking on or adding outcome measures when the patients already have to travel to the site is not so onerous. But when you start thinking about things like gene therapy where the therapy's administered once, it's one and done. I mean, yes, you have to do safety follow-up and assessment to the patient, but the patient and the family no longer has to live close to the site to come and get these biological [unintelligible] infusion center, that sort of thing. So, I'm really intrigued with the explosion of new gene therapy options, and Batten's Disease I guess is one of them. And thinking about how you can, you know, do initial assessments, some at some later point to assess therapy, but what do you do in the interim? Can you incorporate a lot of those remote assessments over time to make it easy for the families?

DR. ADAMS: I mean, I guess I would put that on the table and again just say that it's one option. And if you can reduce burden to -- if you can reduce burden, reduce cost, if offering remote evaluation will enhance participation and the quality of the data that's obtained, and it doesn't compromise any safety for your subjects, and it doesn't compromise the assessment of your outcomes, I don't see why we wouldn't at

least consider it. So, I think, I think it would be on the table, but it stays on the table if the study design can be implemented with rigor. So, I know that's a vague answer. Elsa, did you want to add anything to that?

DR. SHAPIRO: Yes, I do. So I think that in gene therapy trials, that that's a real possibility.

DR. ADAMS: [affirmative]

DR. SHAPIRO: I would say that the way that that would be done is not in the home, necessarily, but in the psychologist's office, in wherever it's being, you know, a site gets selected. And then there's somebody in a central place watching the testing and that the battery is taught to the person there. And then, you know, multiple testings can be done. The other place that I think this might be useful, you know, in our longitudinal study of MPS-1, 2 and 6, we had, we had lots of kids in that study; we ended up with 135 patients in that natural history study. But there were about a dozen patients who were older patients with MPS 6, MPS 2, who couldn't travel because they were too sick. And it did occur to us that we might do that, but it was very expensive for us to set that up in that kind of a natural history study. We didn't have that much funds. But I think that if you were doing a natural history study, and there were patients who were too sick to

travel, but they were older and able to do the tests with the psychologist locally, that that might be a way of gathering natural history data that was appropriate.

DR. ADAMS: Yeah, and I want to jump on that last point because, in the interest of time, I skipped over that on my last content slide. That I think that doing this for individuals who can no longer travel is really an important thing to consider. We recognize that in our evaluations of Batten Disease we've now spent 14 years characterizing the natural history of Batten's Disease, and we've gotten pretty good at characterizing Batten's in the early and mid stages of the disease, but at some point, children are no longer able to travel to our center of excellence; they're no longer able to travel to the family meetings, and so they drop off of our radar. And so we have many fewer numbers of observations of the phenotype in the later stages of the disease. So our understanding of Batten's is restricted to a certain of severity, and by, and so Dr. Augustine who you saw on the small box in one of the slides, she and I have been talking about ways to remotely evaluate, perhaps, some of the later-stage issues. And it may be as straightforward as simply interviewing a parent through video. But the ability to have that direct conversation, and then having the child in the room to observe

the child and lay eyes on the child would be invaluable -- and to lay eyes on them in their home environment when we're not able to travel around the country to see these 130 or so children.

The other thing I wanted to touch on is that Ray Dorsey, who is a telehealth guru at our site nationally, he has been using remote technology for both clinical and for research purposes with Parkinson's disease and other adult disorders. And in Parkinson's disease, what he finds is he can bring patients who have Parkinson's disease to a standardized exam room in the nursing home in which they live. And then he can link with them remotely. This is for clinical, not research purposes. He can link with them remotely as the PD expert, and he can provide care remotely, and it's much better for the family because the spouse doesn't have to bundle their fragile, affected family member, spouse, into the car to drive through two hours of snow in a Rochester winter to get to the visit. So the amount of time involved in a visit goes from 30 or 40 minutes of quality time with your provider, where it used to be four or five hours in the car just to get your 30, 40 minutes of time with your provider, or less perhaps.

And because patients are more satisfied and because they have access now to an expert, their compliance with

treatment is better and therefore their health outcomes are better.

Charlie Richard: Just kind of to follow on that, in another company I worked at we were looking at a severely rapidly progressive neurological disease, Krabbe's globoid cell leukodystrophy, and we working with the families trying to figure out how you can come in. But it was even worse than the MPS's in that the families felt that children were so sick, so we had to plan the whole study. It ultimately wasn't done in natural history study where we get some eager neurologist to cross the country or ready to jump in their car and go out to the family's homes. And, you know, that met with a lot of excitement by the families because you could actually examine the child in the home situation where they felt safe and the felt secure. And, of course, as you might imagine, it's paying for the study and finding people willing to travel to the homes to be able to do that. But I thought that's another kind of alternative. It's not skyping; it's not telemedicine, but it's a way to do it other than asking families to come into the center.

DR. ADAMS: I'll close with one comment, and then I should let other people speak here. But we have one family who we've seen for home visits for a number of years, and we have a

really great relationship with them, and the last time we went for a home visit the family was so comfortable with us being there, and they trusted us so much because we now had been in their home a number of times over the years, that the parents had to run an urgent errand to pick up another kids. And so they left, and they were comfortable with us being in their home alone with their Batten-affected kids while they went off and ran an errand for their other family members. So, that's what, that's the payoff in the long run is the relationship with the families.

DR. PARISI: I have a question about standardizing assessment across sites and whether there's any role for having something we call standardized patients or some sort of way of having the same sort of individual, excuse me, either travel to different sites to just ensure that the robustness of the evaluation is consistent across different sites, or having some sort of a formalized feedback mechanism for the families who participate in assessment, to make sure that all of the assessors at the various sites are consistent in their administration of the tests and the tools.

MS. DELANEY: I agree but I'm going to -- are you speaking of the telemedicine, the assessment, this type of assessment, or --

DR. PARISI: Not necessarily.

MS. DELANEY: Just in general?

DR. PARISI: Yeah, in general

MS. DELANEY: Yeah, so that is something that we have done, going to sites and participating in these. This is what I was describing. I think the most valuable sessions have been the face-to-face interactions with the people who will be doing the assessments. That's where you get the most information about the disease-specific issues that come up when you're trying to do some of these, some of these tests. I don't know about the parent. That's an interesting idea, the parent component and parent feedback. I'm not sure how that would, how you would build that into the study, but it's -- I mean, one of the things we've talked about is we had this correlation on the parent report, or the parent interview on a certain tool that we used, and the actual face-to-face direct testing. That's an argument for -- there's the correlation -- that's an argument for giving the parent information and having this nice correlation. We're doing something, we like to see the correlation because what we're testing directly is consistent with what the parent is telling us.

But yes, I think the best scenario is having someone go to the site, someone with the expertise in the disease and

who's been working with these patients for a long time, to meet with the people who are doing the assessment. It's not always feasible.

DR. PARISI: Do they ever actually do an observation of the person doing the assessment?

MS. DELANEY: Yes.

DR. PARISI: So.

MS. DELANEY: Yeah.

DR. SHAYWITZ: Hi, it's Adam Shaywitz from BioMarin. Question to follow up on the standardization: For a lot of these cognitive tests, as you know, we need to go to many countries, and some of the norms in these countries don't exist. What's your perspective on the need and the feasibility of obtaining norms in multiple countries?

DR. ADAMS: So I think Elsa probably could be up here to answer this also, but I guess what I'll say is that relevant to the discussions in the first session we had today, it's always important to have, I think, an understanding of where your test comes from, what the normative standards are. And where possible, to take a look at how the children in your study, in your disease cohort and population stack up against some normative standard. But at the end of the day, I think for these rare diseases where you have such unique phenotypes, what

maybe more important within the disease is to look within subject change over time, and particularly when you're looking at a trial, using things like age equivalence and developmental quotients, and raw scores allows you to engage in growth-curve modeling, which will look at the within-subject change over time.

And that may be more relevant, so long as you can demonstrate that the constructs you are evaluating are relevant to, are relevant, that so, using FDA speak that you're concept of interest is adequately represented by your clinical outcome assessment. And that that has a logical connection to the disease process and also a logical connection to the intervention that you're evaluating, and at the end of the day that it results in a clinically meaningful change in how the patient feels, functions, and survives.

DR. COMO: I'll just ad from the FDA perspective it's something we wrestle with all the time because, as you've heard today, a lot of these trials require multinational or international trials, and one of the questions we have to think about and ask the sponsors is, "It sounds like your outcome measure is a good one, but, you know, can it be used in Finland?" you know, kind of a thing, so we wrestle with the very same question.

DR. SHAPIRO: I have some thoughts about that, and one thought that I have about that is, you know, it's great to have the translations and then you, and, you know, culturally specific kinds of things, and you go to another country and you -- there are no norms in that country. So, what do you do? So, obviously what Heather suggests is within-child changes. The other thing that can be done is to have a control group of the same age and maybe even a younger group that where you get some sort of age norms. And it doesn't have to be a large, you know, when you norm a test you have, you know, thousands of children. And you don't need that. What you need is a control group. And so, I think that that might be another solution within the country, typically developing children taking that test. So that might be another solution.

MS. DELANEY: I would also add that some of the -- and you could comment, I'm sure, Heather, that the nonverbal measures that we're using are pretty universal. So, what I always say about the, something like the Baileys, I think every child around the world is, maybe they're not stacking nice plastic blocks, but they're stacking something [laughs], all at about the same time. So there is -- I know it's a big concern, but when we look at item -- and this is for the nonverbal items -- when we're looking at those specific set of items, I think, I

think we're pretty comfortable using those across the world in international trial.

DR. SHAPIRO: So, I have a little experience with cerebral malaria and HIV, and we did, I've been working, I had been working with a group who is doing work in Uganda. And they were using the Bailey, the KABC and a couple of tests that we developed for preschool measures of attention and memory. With all of these children with different tribal languages and it worked. They knew, you know, the one thing that we were always amazed at, we had video tapes. We looked at all the video tapes of these kids coming in to do the attention tests because we had to rate them. And one of the things we were amazed at was, this was such a novel experience for these kids that they were very attentive. And so this is -- this is a problem, in a way, because they actually did better than kids did in developed countries because this was a novel experience for them. But, you know, the disease effects were there, and we could sort that out. And so, and that data has actually just been published in the Journal of Infectious Diseases. So, I think that, you know, what Kate says is really true that you can use these tests, especially when they're nonverbal, in many different countries.

MR. RICHARD: [inaudible] push back a little bit there because I know it sounds like the right thing to do. You want

to go to Finland. You get a group of the [inaudible] kids, the normal kids, and you do the normative trial but in the context of the pharmaceutical sponsors, [inaudible] asking a pharma to come and think about how to be able to do this.

DR. SHAPIRO: I wasn't suggesting that. What I was suggesting was a control group of, you know, the same number of kids that you might have in your clinical trial. Just to get a sense of whether, you know, whether the normative material, you know, the scores that you got, were comparable with, in typically developing children in Finland, were they the same as typically developing children in the U.S., for example. That's what I, I'm not saying thousands of kids, no. I think that's unrealistic. That's what test developers do. That's not what, that's not the role of the pharma company. But it is their role to maybe get a control group.

DR. COMO: I think we have time for about one or two more questions.

DR. BARBIER: Well, I just want to offer one last comment on this. This is a discussion that we had with a European cognitive expert from one of our trials, and we brought up exactly this question: Are we really expected to not just translate, validate, but norm this test in, let's say for the sake of arguing, Italian, when we might have one or zero

patients from Italy in this trial? And the experts commented, was, "You know what? We've been using this test in, for the sake of argument, Italy, for two decades, and it's never been normed, and people are using it. It has sort of grandfathered in." That was the first comment, which I thought was very, sort of pragmatic European approach.

And the second point is that if a treatment is supposed to be effective and approvable, the effect of the drug is going to have to be so robust that that little variability that you might introduce by using a term that has been translated and validated but not normed in that particular population, it really is a very small noise in the big therapeutic signal you should have for a drug to be, to be effective enough to get approvable. So I just offer this as some experience we've had in Europe where, of all the concerns about cognitive testing, the approach about the language difficulty, it's there in their daily life anyway. Serbo-Croatia uses the tests that were developed, in [unintelligible]. It is much less of an issue there in Europe, we have experienced.

DR. COMO: Thank you. I have one question for Joanne. How much feedback have you gotten from investigators, and more

directly, push back, now that NIH is moving towards, "You must do this as part of your terms of reward, of award"?

DR. ODENKIRCHEN: So, we've, we were, like, one of the first institutes to actually say we strongly encourage you, and now we say you have to use the CDEs, and it's just the core CDEs. And most people are pretty willing to do it. And what we say is you have to use the core in the highly recommended or explain to us why you're not going to use them. And the way our terms of award are, say that you can't enroll your first participant until we actually review them and sign off on your, on the use of your CDEs. And most people, once they look at them and they understand why we're doing that, and we sit down and review the CDEs with them, most of them are pretty compliant. Over the last two years, we've actually gone through all of our diseases and reduced the number of core CDEs. For example, stroke we started out with 120 core CDEs; now we're down to maybe a dozen or 20. And so most people, with the reduction of the number of core CDEs, most, most researchers are more than willing now to do that.

DR. COMO: There's a strong correlation between money and compliance.

[laughter]

DR. ODENKIRCHEN: Well, well, and also, you know, it's also, if you, now with the whole issue of data sharing, I think people are more -- and we also have a require like most people do, you have to share your data, you know, two years after the last patient was seen or a year after your, your primary publication so…

DR. PARISI: And how are conditions chosen? Is there a nomination process if a group wants to say, "Hey, we'd like to encourage NINDS to do CDEs for X condition”? How is that decision made?

DR. ODENKIRCHEN: We started out with our large diseases. Now, if people want to do it they can come to me and we can talk about it because now we kind of have this hybrid model, and I just, I've been working with them and with the National Library of Medicine in trying to get them going on their own.

DR. BURACCHIO: I think we're ready to wrap up now, so thank you all for the discussion. So, we've approached the end of the day, and so I want to thank all of our speakers, all of our presenters, our chairs today. We really appreciate your time and energy, effort. You've really enlightened us today. And then I also want to give a special thanks to all of our steering committee members, particularly our external steering

committee members, which are Susan Waisbren, Elsa Shapiro, Melissa Parisi, and Gerry Cox, who have volunteered their time and have helped to organize this workshop along with many members of the FDA that Donna Griebel outlined this morning. So thank you very much for your contributions.

I think this has been a really fantastic day. I think we've had great talks across the board, really good discussions, great questions. I couldn't be happier with how the workshop has gone today. And so, I think, you know, just to kind of recap, we went over natural history studies this morning. We had people sharing their experiences with us, expert experiences of scale selection, what has gone wrong, what they would improve the next time around. Hopefully, this is useful advice for many of the companies here who are about to embark on some of these endeavors. We also heard from patients, clinicians about what meaningful outcomes are to them and considerations from the patient perspective about cognitive testings and some of the challenges that the patients themselves face when going through these clinical trials. And in the afternoon, we heard, again, more expert experience on scale development and tips and procedures to go about developing a scale, what to think about when you're selecting your scale or considering doing modifications and again -- and standards in the last session

here, methods for standardizing assessments and for collecting data. So hopefully, everyone found this as useful as I did.

Just to let you know that transcripts will be available online. I'm not sure what the timeline is. A month or two maybe [laughs]? But all the slides will be posted online. If you Google -- I think it's “FDA Neurocognitive Workshop” -- it'll come up. Slides will be posted. Transcripts will be available.

And a reminder, also, that tomorrow, there is another session for those of you who are sticking around for that. Tomorrow will be about long-term neurocognitive outcomes in pediatrics not specific to Inborn Errors of Metabolism but in a more general pediatric population. I believe they'll be touching on epilepsy and oncology for a few things and normal child development. So I think that will be a really great workshop as well. That starts at 8:00 a.m. tomorrow, I believe in the exact same room.

So with that, I thank you. And have a great evening.

[applause]

(Whereupon, at 4:40 p.m., the meeting was adjourned.)

INDEX