Assessing Neurocognitive Outcomes in Inborn Errors of Metabolism

U.S. Food and Drug Administration
Center for Drug Evaluation and Research
Proceedings of Meeting held on April 16, 2015

Table of Contents

Approach to Assessing Cognition and Behavior in Inborn Errors of Metabolism

Tools to Standardize Assessments Across Multi-Site Trials

Workshop Overview

In recent years, a variety of factors have led to development and testing of a significant number of new treatments for inborn errors of metabolism (IEMs). Innovative diagnostics and enhanced newborn screening programs have made it possible to identify IEMs earlier than in the past and begin available treatment sooner. Increased understanding of the pathophysiology of IEMs coupled with advances in genomics that enable scientists to identify variations associated with these diseases have elucidated the mechanisms of IEMs.

Most IEMs begin in infancy or childhood, making it hard to distinguish between changes resulting from treatment versus those that arise from a child's development or from disease progression (Figure 1). Clinical studies must be able to distinguish between these variables to assess treatment effects. Results of natural history studies can help clinicians identify changes associated with a disease that is unfolding in a developing child. Natural history studies should be started early to generate the most rigorous data that can inform treatment approaches.

New treatments hold promise for improving patients' lives. Our ability to accurately measure the effects of treatment on a patient's functioning is essential, yet such assessments pose significant challenges, especially when evaluating neurocognitive outcomes of IEMs. Behavioral and physical symptoms are often difficult to measure and vulnerable to bias. It is therefore crucial that neurocognitive assessments are based on well-defined and reliable measurements that have been validated in adequate and well-controlled studies and can be replicated across clinical settings.

Clinical studies of rare diseases are, by nature, challenging due to limited patient populations. In IEMs, this is further complicated by the many types of diseases that occur; the wide range of ages affected; and the heterogeneity of deficits, which can vary within patients over time as well as between patients. While measures of function must be carefully constructed in order to avoid bias, validity also rests on including input from patients, families, and caregivers when developing these assessments. Meeting these challenges is essential in order to bring safe and effective drugs to market to treat these rare diseases that have a profound impact on patients and their families.

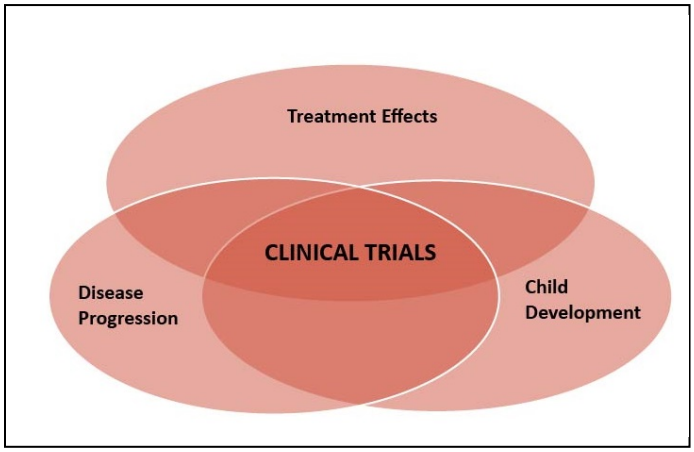

Figure 1. Child development and disease progression intersect with treatment effects in clinical trials. Clinical trials designed to measure treatment outcomes must take into account child development and disease progression -- dynamic factors that are likely to have opposing effects. Standardized, well-defined, and reliable measures of neurocognitive outcomes are essential to enable researchers to assess treatment results reliably. Natural history studies aid in identifying factors associated with disease progression in the context of child development.

Roadmap for Developing Clinical Outcome Assessments for Clinical Trials

Sarrit Kovacs, PhD, Clinical Outcomes Assessment Staff, Office of New Drugs, Center for Drug Evaluation and Research, FDA

Introduction to Patient-Reported Outcomes and Clinical Outcome Assessments

Accurate and reliable measurement of treatment effects in diseases resulting from inborn errors of metabolism (IEMs), based on the development of outcome assessments that are well-defined and reliable, is essential for conducting adequate and well-controlled studies for IEMs. While this section focuses primarily on patient-reported outcome (PRO) measures, it is important to note that with IEMs, many patients are either too young or too impaired to self-report. Observer-reported outcomes, often from parents or caregivers, can be essential.

This section provides an overview of the development of clinical outcome assessments (COAs), which measure a patient's signs and symptoms, overall mental state, or the effects of a disease or condition on how the patient functions. PRO measures are one type of COA and are based on reports that come from the patient about the status of his or her health condition without amendment or interpretation by others including clinicians. As noted above, in IEMs, observer-reported outcomes may be needed, as PROs are not always possible. In this section, we also present potential pathways for stakeholders to interact with the FDA on instrument development. A synopsis of the FDA's COA Drug Development Tool (DDT) Qualification Program, a way to work with the Agency outside of an individual drug development program, is also presented. COAs are qualified specifically within the COA DDT Qualification Program; however there are two other DDT Qualification Programs – Biomarkers and Animal Models. An additional way to communicate with and obtain feedback from the FDA is through Critical Path Innovation Meetings (CPIM). Procedural Guidance can be accessed at: http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM417627.pdf.

The purpose of outcome assessment in drug development is to aid in determining whether or not a drug has been demonstrated to provide treatment benefit to patients. There are three main types of outcome assessments: patient survival, surrogates (often biomarkers), and COAs. To aid COA instrument developers in meeting the FDA's regulatory standard, which states that the methods of assessing a subject's response should be 'well-defined and reliable' (21CFR314.126), the FDA finalized a PRO Guidance for Industry in 2009 (http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM071975.pdf). This provides details on how to establish a PRO instrument's content validity, i.e., the extent to which a PRO measures what it purports to measure in a specific context of use.

Many of the recommendations in the PRO Guidance can be applied to other types of COAs, such as assessments of observer-reported outcomes (including from parents and caregivers), clinician-reported outcomes, and performance outcomes. The PRO Guidance provides an optimal approach to PRO development; however, flexibility and judgment are necessary to meet the practical demands of drug development, such as tight timelines. In addition, the FDA encourages drug sponsors to engage in early and continued communication with the Agency during instrument development and evaluation. To further help instrument developers establish a well-defined and reliable COA, the COA Staff created the diagram, Roadmap to Patient-Focused Outcome Measurement in Clinical Trials, for patient-focused drug development. This can be found on the COA DDT Qualification Program Web page (http://www.fda.gov/drugs/developmentapprovalprocess/drugdevelopmenttoolsqualificationprogram/ucm284077.htm).

It is important that instrument developers first understand the disease or condition of interest before beginning to select or develop a COA. This is especially true with IEMs, where PROs may not always be

possible or may be an option initially but diminish as disease progresses. In these cases, observer-reported outcomes may be needed. Researchers can then identify what aspect of treatment benefit to measure (e.g., how a patient feels or functions in relationship to their health condition or its treatment), and later select or develop a COA.

Developing Clinical Outcome Assessments

There are two ways in which stakeholders can work with FDA to evaluate existing tools or develop a novel COA. The first is the traditional way, which is within an individual drug development program. In this arrangement, the Agency encourages drug sponsors to meet with the medical review divisions early, even in the earliest stages of drug development to discuss selecting or creating a COA.

The second way in which the FDA can work with instrument developers is a newer process, falling outside of an individual drug development program. This way, within the Center for Drug Evaluation and Research (CDER) Drug Development Tool Qualification Programs, includes COAs as well as other types of drug development tools. This process is intended to produce qualified measures for use across multiple development programs. Within the COA Drug Development Tool (DDT) Qualification Program, the agency may work collaboratively with many stakeholders including instrument development consortia, patient groups, individual academic investigators, and drug developers to establish and qualify outcome assessment tools and make them publically available for use across multiple development programs as appropriate. Qualified DDTs are deemed by the FDA "to have a specific interpretation and application in drug development and regulatory review," within the stated context of use. To help explain the qualification process for DDTs, the FDA published guidance for industry and FDA staff in 2014: (http://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm230597.pdf). The COA Staff created the 'Wheel and Spokes' diagram, which represents the process used to develop a COA for qualification and identifies the key components of the documentation submitted for CDER review. See http://www.fda.gov/Drugs/DevelopmentApprovalProcess/DrugDevelopmentToolsQualificationProgram/ucm284077.htm.)

COAs used in clinical trials are not required to be qualified through the COA DDT Qualification Program. However, developing COAs in consultation with the FDA may increase the likelihood that the content and measurement properties of the selected COA will support drug development in important and challenging therapeutic areas for which there is an unmet need for COA development.

Special Considerations for Developing Clinical Outcome Assessments for Rare Diseases

Developing COAs for rare diseases, such as those resulting from inborn errors of metabolism, presents a particular challenge. In these cases, measures of behavior and functioning level may be the most relevant to quality of life and impact on the family. Outcome measures based on normal populations may not be sufficiently sensitive to detect change over time for outcomes that matter most to patients living with IEMs and their families. Instruments designed with sufficient sensitivity to assess outcomes of IEMs will be most useful if they are developed and validated with those patient groups. These are important considerations when developing COAs for all rare diseases. Common data elements (as described later in these proceedings) can provide reliable and valid measures that facilitate sharing of data across sites and performing meta-analyses that allow comparisons of treatment effects across studies. Such research can generate more information about the natural history of diseases resulting from IEMs, lead to better understanding of disease progression, and more reliably assess treatment efficacy.

Natural History Studies in Inborn Errors of Metabolism

Conducting Natural History Studies for Rare Diseases

Jonathan C. Goldsmith, MD, FACP, Associate Director Rare Diseases Program,
Office of New Drugs, Center for Drug Evaluation and Research, FDA

This section reflects the views of the author and should not be construed to represent FDA's views or policies.

Drug development for rare diseases has many challenges including the small size of geographically dispersed disease-affected populations of infants, children, and/or adults; lack of knowledge about the natural history of the diseases; limited or no existing therapies; and the need for development of study endpoints and agreement on the selected endpoints. Understanding the role of patient registries, and designing and conducting natural history studies are essential for developing drugs for rare disorders. To facilitate development of endpoints for trials and to gain acceptance from regulatory authorities, it is important to initiate natural history studies as early as possible in the drug development process. Planning for these studies should begin no later than the proof of principle stage of development, early in the process of phased human trials. Performance of natural history studies should not delay clinical testing for investigational drugs ready to start testing for the treatment of serious diseases with unmet medical needs. While typical non-interventional natural history studies do not usually fall under FDA regulatory authority, to maximize the chances for success, sponsors should seek discussions with clinical investigators and regulatory authorities prior to IND filing. The Rare Diseases Program in the Office of New Drugs/CDER, the Office of Translational Sciences/CDER and the appropriate CDER review divisions offer consultative resources that can assist sponsors with many aspects of drug development. Natural history studies may be especially important in developing potential therapies for treating rare diseases. In exceedingly rare diseases such as inborn errors of metabolism and rare genetic disorders, these studies may be of great value as there is generally, at best, only minimal drug development experience. Clinical knowledge may also be sparse or restricted to a very small number of practitioners or investigators. In addition, diseases of inborn errors of metabolism often manifest early in life including at birth and it can be especially difficult to assess the course of disease in a developing child. Natural history studies can help to distinguish between changes resulting from development, treatment, or the course of disease.

Natural History Studies Versus Patient Registries

The term registry is not synonymous with a natural history study. "A patient registry is an organized system that uses observational study methods to collect uniform data … to evaluate specified outcomes for a population defined by a particular disease…."[1] Registry is a broader term that does not necessarily describe the important characteristics of a natural history study that make it potentially useful for evaluating candidate therapies for rare diseases. Registries may be limited by the use of charts intended for patient care to define and validate common data elements. Communications from sources with varying levels of interest and expertise also limit the value of registries. While registries may be effective (e.g., focused registries can fulfill post-marketing regulatory requirements), they cannot supply the breadth of information provided by natural history studies.

A disease state natural history has been defined as "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."[2] Understanding this definition and its meaning help define the role of natural history studies. These studies are designed with a specific purpose, to improve understanding of specific diseases or one disease. They are prospectively planned and intended to be comprehensive and detailed. They

describe the disease independently of individual observers, clinical sites, and interventions. If properly conducted, the natural history study will provide important information to shape drug development, such as clinically meaningful trial endpoints.

Types and Features of Natural History Studies

There are several types of natural history studies. Retrospective studies may be simpler to perform but they may be incomplete, difficult to interpret, and lack agreed upon medical language. The strongest design is a prospective natural history study, which involves the collection of baseline information of carefully selected and reliably assessed pre-specified measures. This approach is even more robust if it includes a longitudinal follow up of disease-affected individuals. Cross-sectional natural history studies may be useful but they only provide a static, single point in time, perspective of a patient population without a continuing assessment.

The data elements for a natural history study should be selected based on features of the disease under study. Key information to be collected includes factors having the greatest impact on patients and families in terms of how they feel, function, and survive. Potential morbidities and prognostic characteristics need to be captured. Also, disease features over time should be identified to help formulate a sensitive clinical endpoint. Analysis of data from a natural history study can provide information for drug development and facilitate design and conduct of adequate and well-controlled trials. There are challenges in studying IEM populations that natural history studies can address. These studies can describe the full range of disease manifestations and subtypes, including variations in day-to-day severity within patients, and variability between patients (phenotypic differences). To improve precision, a longer duration natural history study may be useful. Therefore, it is important to design and implement the study as early as possible in the drug development process. If there is a change in the standard of care during the study, the change should be implemented and documented, but the study can continue and still provide useful information.

How Natural History Studies Can Help Drug Development

With the findings available from a natural history study, a sponsor can develop and select outcome measures that are more specific or sensitive to changes in the manifestation of the disease resulting in a shortened drug development time. Safety concerns and inadequate efficacy findings in clinical trials can also be detected and addressed sooner to economize development. Data from natural history studies inform the benefit-risk analysis critical to drug development and regulatory decision-making. During the natural history study, the development of new/optimized biomarkers is also possible. These biomarkers may provide proof of concept, guide dose selection, and facilitate early recognition of safety issues. The predictive value of a proposed biomarker, including measurement technologies, can be validated during the natural history study or during the drug development process.

For diseases of inborn errors of metabolism, developing safe and effective therapies presents challenges not generally encountered in developing treatments for more prevalent disorders. Natural history studies, if well-designed and implemented in a timely manner, can enhance and accelerate drug development for these and other rare diseases. These studies should include assessing signs and symptoms most important to patients and families and reflect how they feel, function, and survive. To further enhance the value of natural history studies, developers should make the information widely available to researchers, clinicians, and patients and their families. Data from natural history studies can accelerate the drug development process and availability of new approved therapies shown to be safe and effective.[3]

Challenges of Cognitive Testing in IEM Natural History Studies

Elsa Shapiro, PhD, University of Minnesota

Inborn errors of metabolism are rare congenital defects of single genes that cause disease. Deficient enzyme or transport proteins can cause toxic substances to accumulate, interfere with normal functioning, and reduce the ability to synthesize essential compounds. Although these diseases can present throughout the lifespan, most have their onset in childhood. Most affect the development of the central nervous system (CNS) as well as other organ systems. Many are progressive and result in neurodegeneration with an associated childhood dementia. Assessing cognitive function is therefore an essential outcome measure in both natural history studies and clinical trials.

Historical Background

In the 1980s and 90s, it became clear that one size did not fit all in the cognitive assessment of IEMs. Phenotypes varied by age of onset, severity of disease, and affected organs. For most children with CNS involvement, sensory, cognitive, and behavioral manifestations varied considerably. Initially, natural history data were gathered using clinical assessments whether or not children were eligible for treatment. As time went on, and a broader range of children could be treated, standard protocols were developed to monitor treatment outcomes, but it became clear that standard neurodevelopmental tests and test scores did not always fit the patients, so alternative approaches were necessary.

Without knowledge of the rate of progression, the age of onset, and the severity of each disease, a clinical trial of a treatment would not be possible. It became evident, however, that the rate of cognitive decline could gauge how long a trial needed to run before treatment effects were seen. Because these diseases are rare and the pool of patients is small, it is essential to obtain data from as many patients as possible. Getting the most sensitive and specific data has been the challenge faced since the beginning of research on these diseases.

Dementia in childhood differs from adult dementias as no premorbid history is available for comparison, and the brain is developing throughout childhood, necessitating a 'rate' of development for comparison. Furthermore, children's performance is more sensitive to environmental and medical factors than adults'. Conceptually, the child's performance is the resultant of vectors of normal developmental progress and disease. Using age-equivalent scores instead of IQs allows the detection of slowing, plateauing, or declining abilities that define the resultant. It clarifies the age that these changes occur in each disease. Obviously, a treatment applied before brain damage occurs results in better cognitive outcomes.
Due to the discovery in the 1980s that bone marrow transplantation (BMT) could halt the cognitive decline in Mucopolysaccharidosis (MPS) Type IH (Hurler syndrome), [4] new methods of assessing cognition in these very young children were developed. The spectrum of MPS I ranges from attenuated to severe (Hurler syndrome). Although cognitive abnormalities are found in all phenotypes, Hurler syndrome shows a typical pattern of onset in the latter part of the first year of life, progressing with increasing somatic and neurological involvement. Without treatment, cognition is normal in the first year, slowing of development occurs in the second year, plateauing of development in the third, and loss of skills thereafter, with death by age 10. Early methods determined that the loss of developmental quotient (DQ, which is age equivalent divided by chronological age) was between 15 and 20 points per year. Recently, MPS Type IIIA (Sanfilippo syndrome) was found to have a similar pattern in those patients with the most severe disease. It was also determined in Hurler syndrome that, although BMT halted the disease progression, the trajectories of development after treatment were not normal.[4] Predictors of outcome rate were age at treatment and pre-treatment DQ. The best outcomes occurred in children with normal DQ who were treated before age 2. Understanding the rate of cognitive development in untreated children helped to evaluate the treatment effects in clinical trials.

Guidelines for Cognitive Evaluation in Natural History Studies

From these early studies and later experience with cord blood transplant, enzyme replacement, and other treatments in a number of IEMs, the following guidelines for neurocognitive testing were developed. They apply to natural history studies and clinical trials.[5]

1. **Use a developmental model to provide growth information with appropriate metrics, especially for young and impaired patients.** Pediatricians routinely employ standardized developmental growth curves to assess age-appropriate height and weight gain. Likewise, using standardized developmental growth curves for indicators of cognitive ability, constructed from raw scores or their proxies such as age equivalent scores, permits assessment of whether the child is still developing, plateauing, or losing milestones. Such an approach provides better statistical power with fewer patients due to multiple time points for each child.

2. **Tailor the type of scores to the age and ability of the patient.** Standard scores such as IQ provide a built-in normative comparison and are appropriate for older, mildly to moderately impaired children. In very young children and children with severe cognitive impairment, the floor on these tests renders the results insensitive (most IQ tests have a floor of 50). Studies have found that very low standardized scores are imprecise.[6] An alternative is using norm-referenced age equivalent scores as in the Bayley Scales of Infant Development-III (BSID-III)[7] and the Vineland Adaptive Behavior Scales-II.[8]

3. **Select appropriate tests.**
   a. **Decide if the tests should be disease-specific or use standard generic measures.** Tests need to be chosen that are the most sensitive to change and will yield the most specific knowledge about the disease in a natural history study. General endpoints such as "mental retardation," "cognitive impairment," or "behavior problems" should be eschewed for more specific, detailed information. Preliminary information about the disease and its age of onset (since phenotypes vary with age of onset) should be taken into account when deciding on tests to be used. While generic tests as published are often not useful in rare diseases, adaptation of standard measures can be the most efficient method of gathering data.

   b. **Match the test with the information you have about the phenotype.** Examples: Consider whether the test uses a lot of language-based instruction if language is limited; whether physical handicap interferes with a test requiring motor response; or whether behavioral abnormality will interfere with test performance.

   c. **Assess whether the range of cognitive levels and ages in the disease in question are covered by the test selected.** Are the floor and ceiling of the test disease-appropriate? Will the test cover the entire range of ages and ability levels? A single test that measures the same construct across ages is advisable in any natural history study or clinical trial.

   d. **Select tests that are repeatable.** Frequent testing is acceptable in younger children who, as they develop so quickly, are able to perform new items at each testing. However, some domains, such as problem-solving tests or memory tests, show significant practice effects in older, less impaired children.

   e. **If considering an international trial, translate and norm tests for the countries in question.**

   f. **Other test considerations:**
      - Battery must be **short and focused.** Not all domains need assessment, as they would in a clinical evaluation.

- Tests selected must **answer the relevant questions** and be directed to what is necessary to measure over time.
- The scores must yield an **appropriate metric** useful for tracking change**.**
- Tests selected must have **appropriate normative data**.
- If the study is multi-centered, **quality control is essential**. Training examiners, videotaping or observing examiners, and scoring reliably are important considerations.

g. Behavioral endpoints: Behavioral change can also be an endpoint. Such variables are harder to quantify than cognition, but possibly can be assessed as parent-reported outcomes. Behavioral phenotypes are disease-specific and differ from the range of abnormal behaviors tested by generic measures, thus requiring unique scales.

4. Association of cognitive tests with other measures. It is important to compare cognitive results with other measures in a natural history study, for both clinical and scientific purposes. Such measures might include quantitative ratings of disease progression, quantitative MRI (e.g., based on measurement of brain volume; diffusion tensor imaging that can reveal white matter abnormalities [9], or validated scoring methods), and biomarkers.

5. Challenges encountered in testing young and physically or behaviorally impaired children

a. Sensory and motor problems: Do not give children tasks that they are unable to complete due to sensory and motor limitations. Make sure the child has hearing aids, glasses, and other prostheses.

b. Behavioral difficulties: Investigators and examiners often falsely assume that behavioral difficulties mask real ability. It is more often the case that behavioral problems intensify when tests are too hard. Examiners who test children with behavioral difficulties must be experienced.

c. Fatigue and illness: Do not test when a child has any illness, after a medical procedure, or until 36 hours post-anesthesia.

d. Random variability: There is no way to avoid this except to have the largest sample size (N) possible.

e. Lack of engagement: How do you know if the testing is valid? If possible, when the examiner or parent feels that the child is not performing or lacks motivation, the testing should be discontinued and redone. If motivation is unclear, direct testing such as on the BSID-III can be compared with parent report such as on the Vineland-II. This can only be done after a valid baseline evaluation examines the congruity of performance on these two measures, as then future discrepancies can be gauged based on the initial association of these two measures. Example: In a natural history study of MPS IIIA, both the Vineland-II and the BSID-III were administered and found to be tightly correlated at baseline; discrepancies at a later visit could then be assessed.[5]

6. Requirements for the tester and testing environment: The testing environment should be standardized and child friendly but not distracting; testers must be facile with the test; testers must know the disease and what to expect; testers must be familiar with behavioral problems; and testers must follow the rules but be flexible.

7. **Quality control:** In order to eliminate random variation in the results, testers must have training on the test and the disease; they must be observed or videotaped to establish their facility with the test. Periodic re-observation is recommended. All tests should be re-scored independently.

8. **In order to ensure that patient and family motivation is high and that they will cooperate in a natural history study, the following recommendations are made based on our experience with such studies:**
   - **a.** Employ a coordinator who knows every patient and parent and reaches out to them, spending time with them to explain why we need the information and what the parents and child can get from it.
   - **b.** Explain the testing in detail to parents and the child, if the child can understand.
   - **c.** Include the parent in testing if the child is young, and show the parent the testing room if the child is to separate for testing.
   - **d.** Most importantly, provide feedback in natural history studies regarding how the child is doing – especially in school and psychological issues. This is not always possible in a clinical trial, but it is necessary for parental engagement in a natural history study and provides a valuable benefit for parents.

Conclusions

Measures of cognitive ability in young and impaired children can be reliable and valid if correctly selected and used; test batteries should be focused and short with available normative data. In older, less impaired children, tests need to be repeatable, focused, consistent across ages, and adaptable for handicap. Tests should not be culture-specific. Correct use of test measures requires attention to quality control and examiner training. Investigators should engage parents and communicate why the effort to engage in a natural history study is worthwhile. A recent natural history study demonstrates the use of these recommendations.[10]

New treatments for IEMs are now in trials; others may soon be available for study. Natural history studies, which trace the trajectory of cognitive growth in untreated children, are necessary to identify sensitive and specific cognitive endpoints and the growth rate of these patients prior to clinical trials.

Lessons Learned from Natural History Case Studies

Neurocognitive Outcomes in Urea Cycle Disorders: Lessons learned from the Longitudinal Study for the Urea Cycle Disorders Consortium

Susan E. Waisbren, PhD, Associate Professor of Psychology, Harvard Medical School
Psychologist, Division of Genetics and Genomics, Boston Children's Hospital

The Longitudinal Study for the Urea Cycle Disorders Consortium (UCDC) represents an important example of natural history studies that potentially provide data for development of novel treatments. Over the 10 years that this study has been ongoing, we have taken steps to expand its usefulness to clinicians, patients, and pharmaceutical companies engaged in drug development.

The UCDC is part of the Rare Diseases Clinical Research Network. There are now 14 different centers (with more soon to be added), including 2 in Europe, that form the Consortium. Urea Cycle Disorders (UCDs) interfere with the hepatic ammonia detoxification pathway, leading to hyperammonemia and

other biochemical abnormalities. Children and adults with these conditions are at risk for high ammonia, which is toxic to the brain. Total absence or partial deficiency of the first four enzymes in the urea cycle lead to hyperammonemia, usually within the first few days of life. Deficiency or absence of arginase 1 leads to other neurological and developmental problems, although hyperammonemia is not always present. Ornithine transcarbamylase (OTC) deficiency is the most common of the UCDs and is X-linked, leading to more severely affected males. 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:

Proximal Disorders:

OTC Ornithine transcarbamylase deficiency
CPS1 Carbamoyl phosphate synthetase I deficiency
NAGS N-acetyl glutamate synthetase deficiency

Distal Disorders

ASL (or ALD or ASA) argininosuccinate lyase deficiency/argininosuccinic aciduria
ASS1 (or ASD) Argininosuccinate synthetase deficiency or citrullinemia
ARG1 (or ARG) Arginase deficiency

Transporter Defects

CITR- Citrin deficiency
HHH/ORNT (or HHH) Hyperornithinemia-Hyperammonemia-Homocitrullinuria Syndrome

The rarity of these genetic conditions requires large collaborations, such as the UCDC. This, in turn, presents its own challenges. The four main challenges are: 1) designing and collecting data on diverse populations, 2) capturing data in a uniform way when there are over 14 sites involved, 3) conducting a longitudinal study when novel therapies are introduced along the way, and 4) incorporating appropriate measures when there is a broad range of outcomes.

The UCDC now includes neuropsychological testing data on 562 different individuals with 8 different UCDs plus a few with an unconfirmed UCD. The number of subjects in each group varies dramatically, ranging from 2 with citrin deficiency to 327 with OTC deficiency (Figure 1). In addition, 9 subjects described as “likely to have a urea cycle disorder” had biochemical symptoms of a urea cycle disorder, but confirmatory testing did not lead to identification of a known condition.

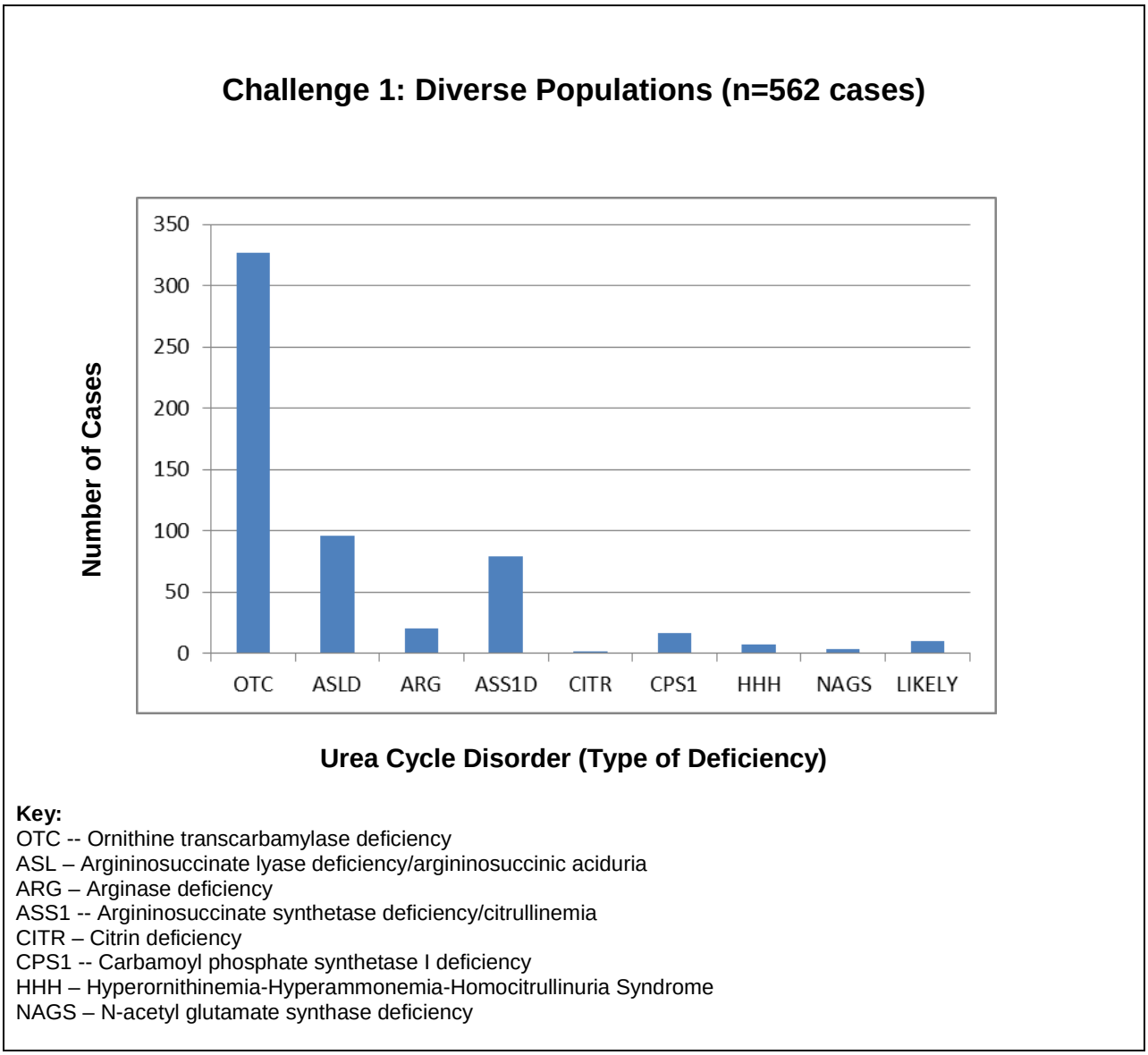

Figure 1. Diverse diagnoses among UCDC study participants. The longitudinal study of the Urea Cycle Disorders Consortium includes participants with different types of UCDs. OTC deficiency is the most common among study participants, followed by ASL deficiency and ASS1 deficiency.

Diversity needs to be considered when examining factors contributing to results from the neuropsychological evaluations (Table 1). For example, many more evaluations were performed on females than males. Age ranged from 6 months to 71 years. Subjects were categorized as being symptomatic (usually indicating having had hyperammonemia, or significant developmental delay or cognitive deterioration). The timing of symptoms and method of identification also varied. Just over 100 evaluations were performed in participants identified via newborn screening and close to 200 were in participants identified because of an older symptomatic sibling or because a younger sibling was identified by newborn screening. The vast majority of evaluations were performed in participants who came to attention because of clinical symptoms.

Challenge 1: Diverse Populations (n=787 evaluations)

CHARACTERISTICS		
SEX	Males: 288	Females: 499
AGE (range: 6 months to 71 yrs)	Infants: 182 Pre-school: 109	School Age: 244 Adults: 252
DISORDERS (Type of Deficiency)	Proximal: OTC, CPS1	Distal: ASS1, ASL, ARG Other: HHH, NAGS, CITR
PHENOTYPE	Symptomatic OTC females: 58 males: 112	Asymptomatic OTC females: 241 males: 0
ONSET	Neonatal onset: 190	Post-neonatal onset: 597
IDENTIFICATION	Newborn screened: 116 Clinically identified: 474	Family history: 193 Unknown: 4

Numbers refer to # of evaluations in database for that category

Key:
OTC -- Ornithine transcarbamylase deficiency
CPS1D -- Carbamoyl phosphate synthetase I deficiency
ASLD – Argininosuccinate lyase deficiency/argininosuccinic aciduria
ASS1D –-- Argininosuccinate synthetase deficiency/citrullinemia
ARG – Arginase deficiency
HHH – Hyperornithinemia-Hyperammonemia-Homocitrullinuria Syndrome
NAGS – N-acetyl glutamate synthase deficiency
CITR – Citrin deficiency

Table 1. Diverse characteristics of patients evaluated in the UCDC longitudinal study.

Diversity in treatments and changes in treatments over the years present the second challenge. Initiation of treatment early in life may have a significant impact on outcome, no matter which of the treatment options is prescribed. Early intervention and special education services may also influence outcomes. And finally, medications for Urea Cycle defects as well as medications for depression, anxiety, ADHD, high blood pressure or other health problems further complicate the picture. In the UCDC study, age at treatment initiation ranged from 1 day-63 years (mean=7.03 $\pm$ 12.50 years). Special education services were accessed by 225 children and 73 children received liver transplantation. There were over 100 different medications prescribed to subjects.

In collaborative, longitudinal, natural history studies one of the most formidable challenges is data capture. The initial neuropsychological testing battery, designed more than 10 years ago, included 47 different psychological tests administered in varying combinations to 6 different age cohorts. This made it very difficult to compare performance as the children got older or to compare younger children to older

children—since different tests were used. In year 8, we revised protocols to improve the quality and quantity of data generated (Table 2). This included changing the initial test battery from 4 hours to a shorter, more uniform assessment battery. We divided the study population into infants, preschoolers, school aged children and adults and increased the frequency of testing to every 2 years. Previously, if a child missed the 8 year-old evaluation, for example, he or she was not tested between the ages of 4 and 15 years. In addition, we recognized the importance of obtaining longitudinal data on all participants using the same neuropsychological instrument. Therefore, we arranged for a self-report or informant report on adaptive behavior to be obtained every 2 years. Finally, the old protocol did not reliably collect pre- and post- liver transplant data. We corrected this as well.

Challenge 3: Data Capture

Total Tests in Protocol (*Initial*) – 47 different tests, 6 age cohorts

Total tests in Protocol (*Revised*) – 21 different tests, 4 age cohorts

INITIAL PROTOCOL	REVISED PROTOCOL	RATIONALE
INFANTS: 2 assessments	6, 12, 24 months	Capture milestones
CHILDREN: 4, 8, 15 years	Every 2 years	Ensure longitudinal data
ADULTS: 1 assessment	Every 2 years for ages 18-26, then once + ABAS-II every 2 years	Ensure longitudinal data
Liver Transplant Cases	No set protocol	Right before transplant, 6 months post, every 2 years

Table 2. Revised protocols implemented in year 8 improved quality and quantity of data and offered more feedback for families.

In addition to the challenge of creating the proper testing battery, there is the challenge of collecting the data. For some tests, such as one of the language tests, only 32 results had been obtained over the 10-year period. For others, such as one of the parent questionnaires, over 500 evaluations were completed. Often, the sites had to rely on a rotating group of psychologists to administer the test battery. Many of these psychologists were interns and most were unfamiliar with UCDs. They often had differing priorities about which tests to administer and some scored the tests in different ways. Before changes were instituted, only 40 percent of participants received more than one evaluation. Some families were reluctant to participate in such a lengthy evaluation, especially since many families never received feedback. Sometimes, the study coordinators presented the testing as an "optional" aspect of the study. And, finally, so many years elapsed between testing sessions that some families simply forgot that they were part of the study protocol. Other limitations noted included failure to obtain information in a routine fashion on autism, anxiety, depression, psychotropic medications, and level of functioning pre-and post-liver transplantation. With the new test protocol, these failures were corrected.

Two instruments were found to be especially useful 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 (ABAS-II)[11], which covers the entire lifespan, from early infancy through adulthood. The ABAS-II is a parent or self-report checklist of a broad range of skill areas related to development, behavior and cognitive abilities. It includes subscales for Communication, Community Use, Functional Academics, Home Living, Health and Safety, Leisure, Self-Care, Self-Direction, Social and Work. Four Composite Scores are derived: The General Adaptive Composite (GAC), Conceptual, Social, and Practical. We recently conducted a validation study of the ABAS-II for identifying individuals at risk.[12] The ABAS-II correlated significantly with direct measures of development using the Bayley Scales of Infant Development and with IQ scores obtained from the age-appropriate Wechsler tests. Moreover, it correctly identified children performing more than a standard deviation below the normative mean on these direct measures 69-74 percent of the time, depending on the age group. For adults, the ABAS-II correctly identified those performing below the cut-off 86 percent of the time. Thus, the ABAS-II can be used to detect infants at risk and serve as a baseline measure for global abilities when direct measurements are unavailable or difficult to obtain. While generally not sensitive to subtle changes in cognitive performance, the ABAS-II could serve as a baseline measure for global abilities.

The PROMIS (Patient Reported Outcome Measurement System) questionnaires are highly recommended by the NIH for use in research studies. These were developed by the NIH and now have extensive datasets available for comparison with our disease groups (see http://www.nihpromis.org/science/PublicationsYears). These are very brief (6-8 questions each) questionnaires that measure self-reported perceptions of functioning and emotional well-being. They provide some of the best measures of quality of life in individuals with chronic health conditions.[13] In the UCDC study, adult questionnaires for anxiety, depression, satisfaction with social roles and activities, emotional/behavioral dyscontrol and cognitive function will be included.

The fourth challenge is variability of outcomes (Figure 2). The mean scores for Developmental Quotient (DQ) obtained from the Bayley Cognitive Composite [7] and IQ obtained from the age appropriate Wechsler IQ Test [14, 15] varied widely between disease groups. For example, among participants with OTC deficiency, mean scores improved while scores either decreased or stayed almost the same for the other disorders. Thus, combining disease groups may be problematic. DQ scores were more consistent with IQ scores for patients with CITR and HHH, but these rare disease groups only included only a few subjects. DQ was not conducted on NAGS participants.

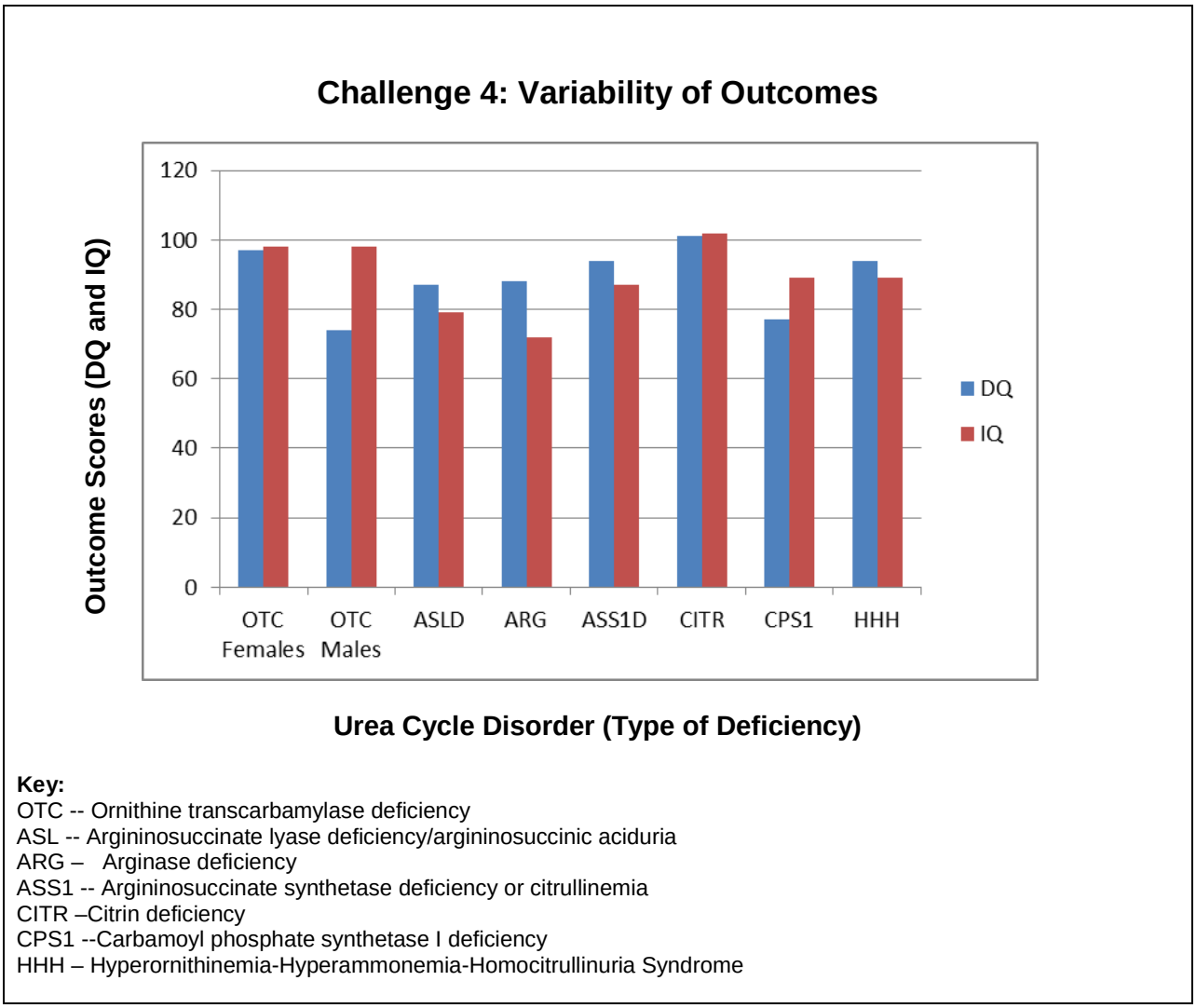

Figure 2. Mean scores for Developmental Quotient (DQ) varied greatly from mean scores for IQ, making it difficult to combine data from disease groups.

Given these challenges, how can data from natural history studies contribute to evidenced-based selection of appropriate outcomes for clinical trials? Data on participants with arginine deficiency provide an example (Table 3). To measure treatment results, we need instruments capable of yielding a broad range of scores that enable us to better distinguish performance differences in ranges typical of participants with UCD.

Using Longitudinal Data to Select Outcomes

Example of Arginine Deficiency (n=20)

Select Instruments with Broad Range of Scores

- **ABAS-II GAC: 40-120**
- **CBCL: Attention 50-80; Total 38-71**
- **IQ: Full Scale 50-111, V IQ 55-107; P IQ 53-114**
  - Blocks 21-60; Matrix Reasoning 20-57; Vocabulary 20-56
- **VMI: 45-113**
- **CVLT: List 1 – 2.5 to 1.5; List 5 – 2.0 to 1.5**
- **Pegboard: -7.48 to 0.49**

Key:
ABAS-II GAC -- Adaptive Behavior Assessment System-Second Edition General Adaptive Composite [11]
CBCL – Child Behavior Checklist [16]
IQ – Intelligence Quotient; VI IQ – Verbal Intelligence Quotient; P IQ – Performance Intelligence Quotient
VMI – Visual-Motor Integration Test [17]
CVLT – California Verbal Learning Test [18, 19]
Pegboard – Lafayette Pegboard Test [20]

Table 3. Score ranges of commonly used instruments for measuring performance in patients with arginine deficiency.

Consistent data revealing change over time would have been very helpful, but among the 20 arginine deficiency cases in the UCDC study, only 3 subjects had evaluations with the same tests on 2 occasions. One case improved, one improved in IQ but declined in visual motor skills, and a third only had a performance IQ which improved.

In summary, the following lessons were learned from this natural history study:

1. When designing a study, focus on research by using more concise instruments that yield measurable outcomes. Our initial very long, diverse protocol resembled an assessment for a child with a suspected learning disability and need for services. For research or clinical trials, the outcomes must be much clearer.

2. In carrying out the study, the entire research team needs to collaborate in developing methods to capture data. The research coordinators and clinicians all need to be on board and recognize the importance of consistent data collection.

3. Data sets need to be accessible. Lack of follow-up or mistakes in scoring of the neuropsychological tests can be more easily recognized if there is a regular review system in place.

4. Collaborative studies are cumbersome: No matter what practices are set in place, large collaborative studies are cumbersome due to changing staff, local populations, inconsistent treatment strategies, and communication challenges.

5. Don't rule out home visits to ensure collection of data.

6. Given how mental health and behavior affect functioning, neuropsychological follow-up studies need to assess for depression, anxiety, ADHD, and autism spectrum disorders. Linking results from these instruments to processing speed may also be important.

In conclusion, natural history, longitudinal studies are not a substitute for piloting instruments prior to a clinical trial. The challenges presented today reinforce the need for a pilot study, which can include a relatively small group of subjects who have the same age, gender, level of baseline functioning, treatment history, and other characteristics as the target population. Moving forward with untested instruments during phase 1 or 2 of a clinical trial can be problematic because invariably the instruments will not be exactly right. Then, a phase 3 trial becomes a pilot study, in which choosing a better instrument is guesswork. On the other hand, a natural history study is perfect for identifying potentially useful instruments for a subpopulation that may benefit from the novel treatment. It can help with grouping subjects according to specific characteristics, such as those with visual motor problems or those with anxiety or low IQ. And, finally, it can highlight the pitfalls that need to be avoided in a clinical trial, such as the consequences of having a very long test battery, depending on testers who may not be fully versed in study techniques due to time constraints or lack of training, or having an overly heterogeneous study group. In the end, a longitudinal, natural history study offers significant rewards for researchers, participants, and industry partners alike.

Natural History Studies in Mucopolysaccharidosis Type II, IIIA and IIIB

Ann J. Barbier, MD, PhD, Shire (former position)[1]

Assessing Abilities of Children with Mucopolysaccharidosis Syndromes

Mucopolysaccharidosis (MPS) is a group of rare genetic diseases with varied clinical presentation. Progressive cognitive decline is a prominent feature of several of the MPS syndromes, including MPSII (Hunter Syndrome) and MPSIIIA and B (Sanfilippo A and B). The diagnosis is typically suspected when developmental milestones of early childhood are missed, or when parents notice loss of acquired skills. The cognitive decline leads to profound intellectual disability. Behavioral issues associated with MPS may interfere with the testing procedures. This makes cognitive testing challenging, as most tests are not designed for such a severely affected population. Our experience in three such populations has identified a number of strategies that can be leveraged to obtain meaningful, longitudinal measurements and reduce the number of invalid or missing data points.

Natural history studies are a valuable tool to help describe and quantify the burden of disease in rare and poorly understood conditions, such as the mucopolysaccharidosis syndromes and other genetic disorders. They can help clinicians identify changes associated with disease and facilitate early intervention, the benefits of which have been well-documented.[21] One of the prominent features of the MPS syndromes is the progressive cognitive impairment, which has until now been well-recognized but poorly described

[1] *Employed by Shire at the time of this workshop*

or quantified. In order to identify potential endpoints for interventional trials for intrathecal enzyme replacement therapy, we conducted three natural history studies in MPSII, IIIA and IIIB. Careful evaluation of available tests, combined with discussions with experts in the field, enabled us to identify tests that could be used in these populations.

These natural history studies had several purposes:

1. To quantify the average rate of decline and variability of the cognitive decline.
2. To investigate brain imaging as a surrogate for cognitive impairment.
3. To investigate a potential correlation between biomarkers in the cerebrospinal fluid and cognitive impairment.
4. To identify any unexpected findings that might lead to specific design challenges in interventional trials.
5. To evaluate the performance of clinical outcome tools that could generate endpoint measurements in an interventional trial.
6. Potentially, to generate data that could serve as a non-concurrent control in diseases where no placebo or other control arm could be included in the interventional trial.

The Sanfilippo A Natural History Study (HGT-SAN-053) [22] enrolled 25 children at a single site, who were followed up to 2 years. The tests included the Vineland Adaptive Behavior Scale, Bayley Scales of Infant Development, and the Kaufman Assessment Battery for Children. The results of the study indicated that in Sanfilippo A, two populations could be distinguished: one which was diagnosed prior to 6 years of age, and one which was diagnosed later than 6 years of age. In the younger, more severely affected patients, it appeared that development was normal up until the age of approximately 30 months, when a plateau was reached, which was then followed by cognitive decline. The patients with MPSIIIA in this natural history study had higher levels of glycosaminoglycans in the cerebrospinal fluid than healthy control subjects, but the levels remained stable within each patient.

Implementing Natural History Studies

Several valuable lessons were learned during the planning, execution and analysis of these natural history studies, the most important of which was that assessments in children with neurodegenerative disorders should be performed by highly trained, experienced assessors. A cognitive assessment performed in a clinical trial setting is very different from one performed for clinical-diagnostic purposes: a balance must be found between implementing a strictly standardized approach versus enabling the child's optimal cooperation by making small accommodations such as allowing breaks, snacks, or the presence of parents in the room. Behavioral interference with cognitive testing must be expected, and testing should not continue if it becomes clear that the numbers generated will not be a relevant measure of the child's cognitive status. We have come to the conclusion that it is better to stop a testing procedure and try again on another occasion, than to persist in the face of significant behavioral challenges, obtain meaningless numbers, and cause stress for the child and family. On some visits, it simply may not be possible to obtain cognitive information from certain children. The concern that repeated testing spaced several months apart might lead to a learning effect and falsely inflated scores does not appear to be of major significance in these natural history studies.

In summary, natural history studies to evaluate cognitive decline are useful tools to help direct the design of interventional trials in diseases of Inborn Errors of Metabolism. It is of the utmost importance to identify clinical sites with appropriately trained and experienced neurodevelopmental psychologists. The choice of the cognitive assessment tools should be carefully evaluated and discussed with experts in the fields, as many of the available tests are not suitable for severely impaired patients. The setup of the study should be sufficiently flexible to accommodate the unique features of these severely affected populations. Psychotropic medications are often used in these populations and whenever possible, dose-adjustments should be avoided during the course of the study. Behavioral issues may interfere with the

testing; in such situations, testing should be suspended and reinitiated at a later time. Whenever possible, sharing results with the families can provide them with useful information and help to maintain motivation for continued participation in longitudinal trials with repeated testing.

What Is Efficacy? Defining a Clinically Meaningful Change

A Caregiver's Perspective on Assessment of Neurocognitive Outcomes in People with Inborn Errors of Metabolism

Melissa Hogan, J.D.

The Importance of Patient and Family Perspectives for IEM Research

The administration and analysis of neurocognitive assessments in inborn errors of metabolism (IEMs) involves the differing perspectives not only of clinicians, assessors, sponsors, and regulators, but also of patients and caregivers (collectively referred to herein as "IEM families"). While the former are involved in the process most commonly as part of their professions, the latter experience the assessments and outcomes as part of their *lives*. As such, their perspectives may be quite different than that of other parties involved in the neurocognitive assessment process.

This section will discuss the issues most important to people with IEMs and their families regarding neurocognitive assessments as well as related factors that can confound efforts to obtain accurate and representative outcomes from neurocognitive assessments in IEMs.

The Unique Roles of IEM Families

One must first acknowledge that the perspectives of IEM families are equally valid, if not more so, than those expressed by other parties involved in the neurocognitive assessment process. Patients and caregivers play an entirely different role in the process than other involved parties – not only are they the subjects of the assessments, but they are also the ones most greatly affected by the results. Some might assume that this fact reduces the credibility of their opinions relative to the otherwise "objective" perspective of the parties designing, implementing, administering, and evaluating the assessments. However, IEM families have the unique personal incentives of effectively treating the patient, accurately evaluating such treatments, and supporting a long-term course of research for their community.
Second, IEM families approach neurocognitive assessments from a very distinct and often dire reality that impacts their view of the assessment process. In many cases, IEMs cause progressive loss of function with a shortened life span. Neurocognitive assessments are either a part of monitoring that decline or attempting to treat it through a clinical trial or otherwise. Neurocognitive assessments are a backbone to gauging the relative success or failure of their efforts to save a loved one, so the incentive to accurately measure is a very personal and highly motivating one.

The assessment methods by which efficacy of experimental therapies are measured, the domains that are selected for measurement, and the assessment process itself are all important issues to IEM families and are vital to IEM families' willingness to participate in neurocognitive assessments and stand by the results thereof.

In measuring the efficacy of experimental therapies via neurocognitive assessments, IEM families fully acknowledge that the first line treatments will not be cures for their or their child's condition. As such, assessments must be sensitive enough to measure a mere slowing of disease progression. In the case of clinical trial endpoints, assessments that utilize age equivalencies, instead of age-normed cognitive scores,

will better measure that slowdown or stabilization and also allow for broader participation in clinical trials for cognitive therapies without sacrificing measurement capability. The distinction between the two is that age equivalencies measure the child's attainment of neurocognitive levels or skills, while age-normed scores measure the progress of the child relative to same-age peers.

For example, in the case of inborn error Mucopolysaccharidosis II, cognition begins to slow and regress between the ages of 2 to 4 years. Many children are not diagnosed until age 4 or after and by that time they are already below an age-normed cognitive threshold often selected for clinical trials. However, the use of age equivalencies would not only more closely track the cognitive loss, stabilization, or growth, it may broaden the potential enrollment population because the focus would be on functional ability, unrelated to chronological age, as opposed to age-relative norms.

Integrating Perspectives of IEM Families is Essential when Planning Studies

As therapies for persons with IEMs someday move closer toward cognitive improvement and even cure potential, measurements should be reevaluated to determine the best method for capturing the reasonable potential of the therapy, be it age equivalencies, achievement tests, normed scores, or alternatives thereto. Selecting the domains to be measured by neurocognitive assessments is also important to the goal of accurately measuring the treatment effect of experimental therapies. While sponsors of clinical research as well as regulators seem to view cognitive ability as the primary proxy for treatment effect of cognitive therapies, to IEM families facing progressive and degenerative disorders, cognitive ability is not the most important endpoint.

In the face of progressive disorders with shortened life spans, cognitive ability might be a possible proxy for disease stabilization, but the domains with the greatest impact are patient behaviors and activities of daily living. As discussed extensively in the 2014 Patient-Focused Drug Development Meeting on Inborn Errors of Metabolism, the aberrant behavioral effects and progressive need for greater assistance with activities of daily living are the greatest challenges to those providing care for people with IEMs. This underscores the importance of observer-reported outcomes (such as by parent/caregivers), clinician-reported outcomes (such as clinician-scored observations of patients in a controlled environment to assess behavior and response to stimuli such as the Risk Room assessments detailed in the later section entitled Sanfilippo Syndrome Behavior Rating Scale), and finally, but possibly the most burdensome to IEM patients and families, performance outcomes (such as neurocognitive testing) to adequately measure these domains.

Finally, the neurocognitive assessment process itself impacts IEM families often on a repetitive basis and thereby, can confound the accuracy and reliability of results. Pre-planning a study or trial to best maintain internal consistency (from patient's first visit to subsequent visits), external consistency (from site to site in a multi-site clinical trial or natural history study), and lateral consistency (from patient to patient) is key in this regard. Involving IEM families or knowledgeable advocates in an iterative planning process is the best means to achieve that objective.

Offering value to IEM families participating in repetitive neurocognitive assessments in the form of assessment results is also key. Children with IEMs often require ongoing assessments for treatment, education, and research purposes, when applicable. Repetitive assessments not only can confound results for each purpose, but can increase medical trauma to children who already endure a substantial burden of medical interventions and therapies.

To reduce stress to patients and their families, the comfort and well-being of the child should be paramount when conducting evaluations. For instance, the evaluator should be experienced and familiar with the disease, what to expect behaviorally, and how best to redirect or address such behaviors in the course of conducting evaluations. The child should be evaluated in a comfortable, child-friendly setting that is optimized with the disease in mind. For example, if a behavioral characteristic is that the child will be easily distracted, the room should be void of toys and other items that might be distracting and require

repetitive redirection. In addition, children should be offered breaks as needed and as testing protocols permit, and have any necessary equipment such as hearing aids or glasses.

Where providing testing results can reduce the overall burden of assessments for each of these purposes, this should be done, since people with IEMs often participate in neurocognitive assessments on a repetitive basis as part of their lives, not just as research "subjects."

Although IEM families have distinct perspectives on aspects of neurocognitive assessments, incorporating these perspectives can be hampered by heterogeneity of disease – cognitively, behaviorally, and physically – which can make assessments challenging both logistically and analytically. In addition, engaging in clinical research or natural history studies (most often the platform for repetitive neurocognitive assessments) is a substantial commitment. Therefore, addressing these issues in the planning stages and integrating perspectives of knowledgeable IEM patient advocates and families, can help reduce this burden, potentially supporting enrollment, retention, and reliability of outcomes.

However, I note the unique situation created when neurocognitive ranges act as inclusion criteria for experimental cognitive therapies with the first-line potential to slow the progression or stabilize an IEM condition. In these cases, consistent with the natural human instinct to protect or save our young, some IEM parents may engage in strategies to increase the likelihood of their child's qualification. Parents know they have a narrow range of time to potentially save their child's life, as the disease progression between a study's enrollment and possible FDA approval of that therapy will likely result in significant skill loss or loss of life. They subscribe to Eminem's philosophy: "You don't get another chance. Life's no Nintendo game." Although neurocognitive inclusion criteria may be necessary in initial trials to determine efficacy, sponsors must be aware and understand that parents may naturally attempt to influence their child's performance on direct measurement or alter their report of the child's skill levels on parent-reported measures. As such, utilizing inclusion criteria other than neurocognitive scores (such as biomarkers), or utilizing broad neurocognitive criteria via age equivalencies could avoid this dilemma. If strict neurocognitive criteria are required to establish efficacy, after efficacy is established, entry criteria may be broadened, diminishing this challenge.

As in most areas of IEM research and treatment strategies, patients are the key to unlocking the most effective and efficient neurocognitive assessment outcomes. Doctors may know the disease medically, pharmaceutical company sponsors may know the therapies pharmaceutically, but only people with IEMs and their caregivers know the life, including the abilities and challenges, the disease in action, and the therapies in action.

As such, their perspectives on assessment methods and domain measures, as well as the assessment process itself, should be incorporated from an early stage prior to the start of clinical research or natural history studies. An integrated and iterative process involving IEM advocates and families can best address the challenges of disease heterogeneity, significant clinical research commitments, and meeting the ultimate goal of saving the lives of people with IEMs.

A Clinician's Perspective

Jonathan W. Mink, MD, PhD

Clinicians in a variety of specialties care for people with inborn errors of metabolism. For disorders accompanied by physical disability, pediatric neurologists collaborate with specialists in physical and rehabilitation medicine and with a variety of therapists. Clinicians have traditionally been trained to focus on the pathophysiology of disease -- what's wrong with cells or chemical pathways, the resulting impairments, and laboratory measures associated with these changes that can predict the course of disease. In IEMs, such numbers are often not available and, when they are available, may not predict outcomes. While biochemical and other quantitative measures are important and contribute significantly to research on IEMs, the most meaningful measures for individual patients and their families often reflect the ability to function and to participate in society. Existing rating scales measure behavioral and physical changes. Yet, they are often limited in their ability to predict the impact a change makes on an individual's quality of life.

Many disorders resulting from IEMs are progressive and lead to premature death. Although our goal is to completely reverse the effects of IEMs and restore all normal function, we have almost no disease resulting from an IEM for which we can do that today. Despite the recent growth in research on experimental therapeutics for these disorders, the ability to restore function eludes us and halting disease progression remains the most realistic goal. To date, most interventions offer only stabilization with occasional potential for mild reversal.

In slowly progressive diseases, it can be difficult for clinicians and even parents to assess the degree of change that a patient is experiencing. Scales exist to evaluate the severity of disease, but don't always measure what matters most to an individual patient. Parkinson's disease for instance, has one of the most reliable and widely used scales of any neurological disorder, the Unified Parkinson Disease Rating Scale (UPDRS). That scale includes assessment of tremor in every limb and in different settings. Yet for many people with Parkinson's disease, the tremor is not the factor that most restricts their ability to function. Rather, it's the falling, freezing, and inability to walk independently that are often the sources of the most limiting disabilities. They are under-represented in the UPDRS. Thus, even well-validated, highly reliable, and widely used scales may have limited ability to measure changes that are most meaningful to the majority of patients.

Many years ago, I treated a 16-year-old girl with cerebral palsy. She had a beautiful motorized wheelchair that she could not drive because her movement disorder was so severe and she had so much shaking. Nor could she hold a cup to drink. In fact, she could do almost nothing for herself. Her cognition was quite normal but her communication was severely impaired. Treatment yielded a very small increase in measures of motor function. This small improvement, however, enabled her to use one arm to drive her wheelchair, and hold a cup of water with a top and a straw so she could drink independently. Having the ability to navigate using her wheelchair and to decide when to drink from that cup enabled her to enjoy a new level of freedom, yet was barely discernable on a rating scale.

Those of us who are involved in translational research bridge clinical research and clinical care. In both capacities, we must understand the unique and varied perspectives of patients with IEMs and their families. It is from them that we learn what makes the biggest difference in everyday life. A clinician may document small areas of improvement in a child, yet an important concern may not be ameliorated. For instance, a child may show improvement in ability to sit up, but if anxiety, perseveration, and aggressive behavior are no better, the impact on the functioning of the child and the family may be much less than expected.

My perspective on outcome measures and how to judge the effectiveness of treatment is that the most significant outcomes are those that maximize quality of life by facilitating independence and participation in society. As a clinician, working with patients and families to fully understand what has the greatest impact on their everyday lives establishes the most effective partnership and is likely to yield the most meaningful clinical and research outcomes.

Assessments in Early Interventions for Presymptomatic Disease

Florian Eichler, MD

What is the Significance of Identifying Presymptomatic Disease Associated with Inborn Errors of Metabolism?

If we found early specific and sensitive signs, this could dramatically change how we treat patients with inborn errors of metabolism (IEMs). We could have an earlier window of intervention. We could tailor our treatment towards specific phenotypes because of better understanding how the disease evolves and the associated phenotypes. We could have earlier measures that we could follow over time, and ultimately more successful trials as we often are not treating early, more reversible disease.

What Are the Challenges in Assessing Presymptomatic Disease in Children?

One of the challenges is that the time of attaining milestones is quite variable and each child develops skills at a slightly different time point. Most importantly, the structures of children's brains are 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, differentiating developmental delay from regression is challenging.

To illustrate the dilemma in differentiating delay and loss of skills, let me present disorders of myelination of the brain. Myelin represents the insulation of the wiring in the brain that facilitates rapid conductance of signals across an axon. An important scientific insight was that around the time of birth there is very little myelin in the brain. Brain myelination largely develops in the first years of postnatal life. It starts from the peripheral nervous system and moves into the central nervous system, starts from the sensory tracts and moves into the motor tracts, from the back of the brain into the frontal regions of the brain. So if you look at a child's brain MRI at birth there is little myelin, and hence presymptomatic myelination status may reflect development not disease.

So our success in evaluating cognition and development is much better once myelination is complete. Hence, we are in a better position to assess presymptomatic disease for disorders of metabolism that strike once myelination is complete. An example of this is adrenoleukodystrophy (ALD). This monogenic disorder is due to mutations in the *ABCD1*gene leading to elevations in very long-chain fatty acids and a wide range of phenotypes ranging from isolated dysfunction of the adrenal gland, to inflammatory demyelination in the brain and, in adults, a non-inflammatory axonopathy of the spinal cord.

Importantly, there is one gene responsible for ALD, but several phenotypes. This phenotypic variability is still poorly understood and an enormous challenge to assessing the presymptomatic disease state. Within one family you can have the severe childhood form of brain inflammation followed by rapid decline and death, but also patients with normal appearing brain but spinal cord degeneration causing bladder and bowel problems.

In inborn errors of metabolism, presymptomatic assessments have long focused on chemical biomarkers. This is quite attractive as there is often a single gene that interrupts a single enzyme and leads to build-up

of certain substrates: sulfatides in metachromatic leukodystrophy, very long chain fatty acids in ALD, and N-acetylaspartate in Canavan disease, to name only a few.

However, chemical biomarkers may reflect the path of physiology but do not predict the disease course and that is the big challenge. So, very long chain fatty acids found in plasma are an excellent biomarker for diagnostic purposes -- they are easy to measure, highly specific, and detectable. However, the level of chemical abnormality in plasma does not reflect severity of disease or predict phenotype. This remains a major obstacle in assessing inborn errors of metabolism.

Imaging as a Quantifiable Measure of Brain Disease Burden

In ALD, brain MRI provides much more sensitive and specific indications of early changes than neurologic or neurocognitive assessments. Brain lesions on MRI correlate exquisitely well with survival in ALD. Patients who have no lesion in the brain have a better prognosis for survival than patients with a larger lesion, whose survival is typically much poorer. So the presymptomatic phase in ALD can largely be followed by imaging alone.

Another reason that early assessment is important is that successful interventions often depend on early treatment. For the most part there is a strong correlation between the age of symptom onset and severity of progression, informing our decisions on when to treat. Hence, identifying presymptomatic disease facilitates rapid early recruitment and should allow us to treat more successfully in the early stages. If you employ hematopoietic stem cell transplantation in boys with cerebral adrenoleukodystrophy who have small and early lesions, their chance for survival is much improved compared to boys with advanced disease.

So, brain MRI has helped us define the optimal window of intervention, and advanced imaging promises to bring further progress. In the case of ALD, contrast enhancement accurately detects more rapid lesion progression. Recently we have gone further to look at MR perfusion, an advanced imaging tool that shows passage of blood through the brain's vascular network. We find that normal appearing white matter beyond the lesion edge -- areas not yet affected but where pathology is about to ensue -- show low cerebral blood volume. We have used this to help predict whether patients are likely to have lesion evolution. Also we have seen that once bone marrow transplant successfully halts progression, contrast enhancement disappears and the lesion stabilizes, as shown by the MR perfusion.

Importantly, disease progression is not the only reason to treat early. We must also take into account that the intervention itself takes time. 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 assert beneficial effects as mononuclear phagocytic cells. 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 protein in the peripheral cells and in these early symptomatic stages we can watch levels of very long chain fatty acids go down, but truly meaningful are the changes in the brain, the impact upon the evolution of the lesion, and the corresponding effects on the patient.

Optimal Presymptomatic Assessments

So what are the optimal presymptomatic assessments in neurological Inborn Errors of Metabolism? The optimal presymptomatic assessment depends on what part of the nervous system is affected by the disease. So if the brain is affected in leukodystrophy, that is where you want to look. On the other hand, if it is the alpha motor neuron and the spinal cord, then you want to use different assessments. If it is the peripheral nerve, you want to carefully look at nerve physiology.

In hereditary sensory neuropathy type 1, we have been able to develop a very useful biomarker by examining the intraepidermal nerve fibers (IENF). This inherited neuropathy affects the distal limbs of arms and legs. Punch biopsies in the thigh and in the distal leg allow us to directly look at the nerve

fibers. We found that most of the patients lost some IENF in the thigh and that even in the presymptomatic stage there are no IENFs in the distal leg.

Patient-Centered Research

Finally, a word about patient-centered outcomes and our efforts to create a consortium around patient-centered research called ALD Connect. We were able to bring together many stakeholders in the field, including patients, patient advocates, scientists, physicians, and industry representatives. Through this inclusive endeavor, we were able to grow rapidly, and within a year became one of the 18 patient-powered research networks funded by the Patient-Centered Outcomes Research Institute (www.pcori.org). Here we have had a chance to gain insights into what is important to patients at the presymptomatic stages. I often find parents keep thorough records of their children, and much can be learned about their social milestones crucial to the presymptomatic assessments. Through patient learning academies, patient portals, and Web sites, we are beginning to query this information. Because patients and their families offer unique perspectives, gaining input from them as early as possible is essential for successful research on IEMs.

Presymptomatic Assessments have Potential to Transform Disease Outcomes

Chemical biomarkers and inborn errors of metabolism are useful but not necessarily predictive of disease course. Variability in development and cognitive milestones is a great challenge in using neurocognition as presymptomatic markers of disease. 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 for the specific diseases occur. So structural vulnerability, the selective vulnerability in inborn errors of metabolism is what determines the best assessment. For instance, brain MRI and perfusion may be most useful for cerebral ALD; for metachromatic leukodystrophy you might include nerve conduction; in spinal muscular atrophy, you might use nerve physiology. In hereditary sensory neuropathy, you may want to directly evaluate the peripheral nerve and the small fibers in the skin. So it really depends on the disease. In the coming years, we will likely see that patient-reported outcome will enhance neurocognitive assessments in the presymptomatic stages. Early intervention and presymptomatic disease can be transformative. In particular, our community has been quite galvanized by newborn screening coming about, by establishing a consortium, and by new treatments that are changing the lives of patients.

Approach to Assessing Cognition and Behavior in Inborn Errors of Metabolism

When Do We Need Disease-Specific Scales? How Do We Develop Them? An Example from Glucose Transporter Type 1 Deficiency Syndrome

Alison Skrinar, PhD, MPH, Ultragenyx Pharmaceutical Inc.

Selecting or Designing Instruments to Measure Concepts of Interest

The combination of small groups of patients with heterogeneous manifestations and limited aggregated clinical data on disease burden poses significant challenges to drug development for orphan diseases. Understanding the targeted patient population is critical to selecting meaningful clinical outcome measures to evaluate the natural history of a disease and establish the efficacy of an investigational product in a clinical trial setting. Systematic literature reviews and patient-focused research, including interviews with patients and caregivers, clinical observation of patients, and pilot testing of measures, play a critical role in the clinical trial design process. Results of qualitative research initiatives can inform the selection of endpoints by exploring the potential to utilize established instruments that measure some or all of the concepts of interest or identifying the need to develop a disease-specific instrument that

focuses on the target population. Either approach should result in the selection of assessments that are relevant, can be reliably and safely completed by the subject, and are sensitive enough to characterize the impairment and to detect change in response to an intervention. These elements are particularly important for the characterization of rare disorders and evaluating the efficacy of potential therapeutics to treat these debilitating diseases.

Characteristics of Glucose Transporter Type 1 Deficiency Syndrome

Glucose Transporter Type 1 Deficiency Syndrome (Glut1 DS, also known as De Vivo disease) is a rare genetic metabolic disorder caused by mutations in the *SLC2A1* gene and is characterized by a deficiency of the Glut1 protein that results in the impaired transport of glucose to the brain across the blood-brain barrier (BBB).[23] Glucose is the primary energy source of the brain and insufficient levels lead to significant impairment of brain function and development.[23] The clinical presentation of Glut1 DS is often heterogeneous with respect to the frequency and severity of symptoms experienced by patients.[24] The classic phenotype of Glut1 DS, which affects approximately 90 percent of individuals with the condition, is characterized by infantile-onset epileptic seizures of varying types, including generalized tonic-clonic, focal, myoclonic, atypical absence, atonic, and unclassified. Patients with the classic Glut1 DS phenotype can also experience complex movement disorders, as well as cognitive impairment ranging from mild learning difficulties to severe intellectual disability, delays in achieving developmental milestones, and speech and language impairments.[25, 26] There is no gold standard for assessing cognition in Glut1 DS, and expressive and receptive language deficits complicate the administration of many existing measures. A tailored and flexible cognitive test battery is needed to evaluate cognitive function in patients with Glut1 DS.

Assessing Cognitive Function in People with Glut1 DS

The Cambridge Neuropsychological Test Automated Battery (CANTAB) (Cambridge Cognition, UK) provides an assessment of cognitive function across several domains and neural systems. The test accommodates a broad spectrum of cognitive abilities and can be administered to children and adults. The test does not rely on language skills, which makes it ideal for patients with language disorders and feasible for use in global studies where language barriers can be a challenge. Pilot testing was conducted with patients with a wide range of impairment to identify the subtests that were most appropriate for use in patients with Glut1 DS given the cognitive presentation of the disease. The testing was performed to advise the design of a clinical trial for an investigational product. The four subtests chosen were Paired Associates Learning (PAL), Reaction Time (RTI), Spatial Span (SSP), and Spatial Working Memory (SWM). The PAL test provides a visual memory assessment of episodic memory and new learning that is most sensitive to changes in medial temporal lobe function. The RTI test is an attention test that measures movement and speed of response to a visual target. The SSP test is a measure of executive function, working memory, and planning that is most sensitive to changes in frontal lobe function. The SWM is also a test of working memory that evaluates the ability to retain visual information and manipulate it in working memory using heuristic strategy. The order of the test administration is designed to gradually increase the level of difficulty to prevent discouragement and fatigue. Each test can be aborted at the administrator's discretion and some tests will "time out" based on the number of errors. Normative values are available for each of the CANTAB subtests and allow for the comparison of a subject's ability to an age-matched control group.

Customizing Testing for Glut 1 DS

No consistent measures of cognitive ability for Glut1 DS have been reported in the literature. Language deficits in Glut1 DS also complicate the administration of standard cognitive batteries. The complex nature of the cognitive issues in Glut1 DS highlighted the need for a customized battery for use in Glut1 DS. CANTAB is a nonverbal battery of tests that can be customized to evaluate disease-specific areas of impairment such as mental and motor speed, attention, episodic memory and executive function. A

customized version of the CANTAB system is currently being used in a phase 2 clinical study of triheptanoin therapy for Glut1 DS to evaluate the potential effect on cognitive function in this patient population.

The Unified Batten Disease Rating Scale: A Multi-Axis Scale for a Rare Childhood Neurodegenerative Disease

Jonathan W. Mink, MD, PhD

Forms and Features of Batten Disease

Batten Disease (neuronal ceroid lipofuscinosis) represents a group of disorders characterized by neurodegeneration and intracellular accumulation of an auto-fluorescent lipopigment. Together, they represent the most prevalent class of childhood neurodegenerative disease. The neuronal ceroid lipofuscinoses encompass several distinct biological entities that vary in age of onset, specific neurological phenotype, and rate of progression.

At least 10 forms of Batten disease have been described. These disorders share some important features including vision loss, epilepsy, a movement disorder, and progressive dementia. However, they are biochemically distinct and the specific symptoms and rate of progression differ across forms. It is increasingly clear that knowing the causative gene and mutation does not always predict phenotype either within or across genes. Furthermore, age at onset of symptoms and specific phenotype does not necessarily predict genotype.

Because of the different clinical and biochemical features of the various forms of Batten Disease, it has been clear that each form must be considered as a different disease for the purposes of characterizing the natural history, developing treatments, and assessing the impact of potential treatments. To that end, we developed the Unified Batten Disease Rating Scale.

Developing and Testing the Unified Batten Disease Rating Scale

CLN3 disease (Juvenile Batten Disease; Juvenile NCL), is one of the more slowly progressing forms of Batten disease, with onset of symptoms typically occurring between age 5 and 7 years with progression to death in the third decade of life. Although the gene is known, the function of the protein product is not known and there are currently no valid biomarkers. Thus, characterization and quantification of the natural history was imperative to provide clinical baseline and outcome measures for future clinical trials.

The UBDRS was initially drafted based on a review of the existing literature on different forms of Batten Disease, with an emphasis on CLN3 disease. The UBDRS includes information on subject demographics, previous diagnostic testing, medical history, and medication taken. The subscales include: 1) Physical Assessment, assessing vision and movement with 20 items leading to a severity score between 0 and 112; 2) Seizure Assessment, assessing seizure types, frequency, and impact with 12 items leading to a severity score between 0 and 54; 3) Behavioral Assessment, assessing the frequency and severity of mood and behavior symptoms with 10 items leading to a score between 0 and 55; 4) Capability Assessment, assessing function participation in school and at home with 10 items. In addition, the presence and approximate date of symptom onset is recorded for difficulties with vision, motor, behavior, cognition, seizures, feeding, and sleep. This allows for estimates of the sequence of symptom onset and times between the onsets of specific symptoms. Finally, a clinical global impression is recorded for the key symptoms domains and for overall symptom severity.

The UBDRS was field tested on 23 individuals with clinical diagnoses of Batten Disease, most of whom had the juvenile form. Items were added, modified, or eliminated based on that testing. Subsequently, items were eliminated or modified based on formal reliability testing of the individual items and of the

subscales. [27] Continued assessment of scale performance and reliability with modifications as guided by the data was done until 2007, when the current form of the UBDRS was finalized.

Rare diseases pose substantial challenges to subject ascertainment and recruitment. To mitigate some of those challenges, we took several approaches. We established a contact registry in 2001 that allows us to contact individuals in the registry to participate in current and future research. We attend the Annual Meeting of the Batten Disease Support and Research Association (BDSRA) to recruit participants and to assess affected individuals. [28] We have attended that meeting every year since 2002. We established a Batten Center at the University of Rochester in 2005 (URBC), to recruit individuals who do not attend the BDSRA. The URBC has been named a Batten Disease Center of Excellence by the BDSRA, which has further enhanced our ability to recruit. Finally, we genotype all subjects seen, in order to confirm the specific mutations and to allow assessment of genotype-phenotype associations.

We have demonstrated validity and reliability of the UBDRS in multiple domains. [27, 29-33] Further, we have shown that the UBDRS can be administered reliably using telemedicine.[34] In addition to characterizing the natural history of CLN3 Batten Disease, we have been able to use the natural history data to test specific hypotheses about the disease and have found: 1) females have a more severe disease course than males with a later onset and earlier loss of functional independence and earlier death than do males [35]; 2) the medication flupirtine, which had been shown to have some anti-apoptotic effects *in vitro*, does not appear to affect disease progression in children with CLN3 disease.[36] The UBDRS is currently in use as a secondary outcome measure of an FDA-funded Phase 2 trial of mycophenolate mofetil in CLN3 disease (clinicaltrials.gov NCT01399047).

Current Usage of the Unified Batten Disease Rating Scale

The UBDRS is a valid, reliable, and utile clinically rating scale for Batten Disease. It has provided important natural history data for characterizing CLN3 disease, and is a meaningful potential outcome measure for future clinical trials. It can be administered remotely, which enhances the potential for assessment with reduced burden of participation on research subjects and their families. It has performed well in both generating and testing hypotheses. Although it was designed for Batten Disease specifically, it has been used as a model for developing a rating scale for Wolfram Disease, another rare neurological disorder in children [37] and is a potential model for natural history research in other rare neurodegenerative diseases in children. Ongoing research is focused on measuring the UBDRS response to interventions and predictive validity.

Sanfilippo Syndrome Behavior Rating Scale: Steps in Developing a Disease-Specific Measure

Elsa Shapiro, PhD, University of Minnesota

Background

While standard measures of *cognition* can be used in various diseases and ages, standard measures of *behavior* are more difficult to use in inborn errors of metabolism (IEMs). Usually, behaviors associated with IEMs are related to either specific brain involvement, or the patient's reaction to the physical and cognitive manifestations of the disease which are not disease-specific and can be assessed with generic measures. Examples of disease-specific behavioral problems can be seen in the attention problems that are associated with white matter abnormalities in mucopolysaccharidosis types I and II, the anxiety seen in Batten disease, or – in the case presented here – the severe and unique behavioral abnormalities observed in Sanfilippo syndrome Type A. Establishing disease-specific measures can benefit patients and advance understanding of IEMs. Such measures can aid in diagnosis; help to identify root causes of aberrant behaviors; determine the most effective treatments; improve quality of life for patients and their families; and enhance the quality of research.

The literature on Sanfilippo syndrome has reported consistently that patients have severe behaviors that are challenging for parents and teachers and that interfere with quality of life.[38, 39] Patients are described as heedless of danger, and parents must expend considerable effort to keep them safe. They have also been described as lacking empathy and social skills, and as having autistic-like behaviors. We hypothesized that understanding the brain abnormalities associated with Sanfilippo syndrome would shed light on the neurobehavioral pathology and vice versa. To explore this hypothesis required establishing a standardized, reliable scale to accurately measure behaviors associated with Sanfilippo syndrome, which would allow correlation with direct measures (such as by MRI) of regional changes in brain structure or function.

Here is a step-by-step summary of what needs to be done to establish a disease-specific behavior rating scale and how we carried it out with Sanfilippo syndrome. The steps we took can be applied to other diseases.

Step 1. **Read the descriptive clinical literature and assess the significance of the behavior associated with the disease.** In Sanfilippo syndrome, most articles in the clinical literature made note of the behavioral abnormalities [38, 39] and described them as severe, but detailed descriptions were lacking. We read the charts of patients seen previously in our clinics and obtained descriptions of the behaviors from parents. We determined that the behavior was often very severe, qualitatively unusual, and interfered with parenting and family activities to a significant degree.

Step 2. **Get an expert opinion with observation either in person or on video.**
We obtained the opinion of a colleague who was expert in aggressive and disruptive behavior and had him observe children in a free play situation. We also asked parents to provide detailed descriptions of the behaviors that were most disruptive. A few of their videos were also examined. Our expert noted the consistency of their behaviors, and identified these behaviors as typical of Klüver-Bucy syndrome (K-Bs). (The signs of K-Bs, including orality (repeated mouthing of objects), lack of social reciprocity, and diminished fear, have been shown to be associated with amygdala dysfunction in many species including non-human primates.[40]

Step 3. **Form and test hypotheses.** Children with Sanfilippo syndrome exhibited aspects of both Klüver-Bucy-like syndrome and autistic-like behaviors. We therefore hypothesized that they would meet the criteria for both diagnoses, and devised studies to test both hypotheses. Testing whether Sanfilippo

syndrome patients met the criteria of K-Bs was carried out in a behavioral laboratory "Risk Room" situation modified from the Laboratory-Temperament Assessment Battery.[41] This was done with consent of the parents and the presence of one parent in the room. The test assessed the social interactions, fearlessness, and startle reactions of affected children through staged encounters with people, objects and loud noises. Testing for autistic-like behaviors was done via a standard diagnostic measure, the Autism Diagnostic Observation Schedule.[42] We confirmed both hypotheses and published those results.[43, 44]

Step 4. **Generate items for a scale.** We created items for a scale—the Sanfilippo Behavior Rating Scale (SBRS)—based on theory, lab and clinical results, observation, and parental reports. The requirements for the items were: 1) that they were behaviors typically observed in children with Sanfilippo syndrome; 2) they were concrete and detailed; and 3) parents could understand them. Items were scaled based on frequency, not severity, as judgments of severity are less reliable. Using frequencies such as "occurs all the time," "many times," "some of the time," or "never," accounts for normal children occasionally demonstrating these behaviors. To gather additional data, we added questions related to the ages at which these signs onset and/or disappeared.

Step 5. **Classify items on the scale.** We organized our items into domains such as movement abnormalities, diminished fear, mood, anger and aggression, social abnormalities, orality, masturbation, attention problems, and hyperactivity, all of which we have noted in observations and are associated with K-Bs and autism.

Step 6. **Prune items by frequency of endorsement.** To determine whether these items were characteristic of Sanfilippo syndrome, we gave the 73-item scale to the MPS Society to send anonymously to 100 parents of Sanfilippo syndrome patients. Based on 47 responses received, we could immediately eliminate 5 items that were infrequently endorsed. We also asked for parent volunteers for a telephone interview to review the items, and carried out open-ended interviews to explore whether parental descriptions concurred with the item endorsements. The resulting scale consisted of 68 items.

Step 7. **Carry out reliability studies and cluster development.** Cronbach's alpha was calculated for all items to measure the internal consistency of the entire SBRS scale and the clusters. We examined the associations of each domain with the predefined overall clusters: Movement, Lack of Fear, Social-Emotional Abnormalities, and Executive Functions. Orality and Mood/Anger/Aggression became stand-alone domains as they did not correlate with any other items. We then created Loess scores, a method to examine the trajectory of mean scores across the age range. From these analyses, we are now able to score the SBRS on these four clusters and two domains. Although we did not do test-retest reliability studies at that time because of insufficient numbers of patients, we are now in the process of carrying out such studies.

Step 8. **Carry out validation studies.** We examined the concurrent relationship of the SBRS to the Autism Diagnostic Observation Schedule (ADOS) scores, the Vineland Adaptive Behavior Scales II,[8] measured outcomes from the Risk Room, and manually traced amygdala volumes.[10, 22] As an indication of concurrent validity, we confirmed that SBRS individual clusters were associated with these external measures. We then did cross-validation with a small sample of Sanfilippo syndrome type B patients, and found similar results (except for differing age of onset of signs) from retrospective parental reports.

Step 10. **Publish manuscript and make the scale available to practitioners and researchers**. This has been accomplished (see references).

Step 11. **What we haven't done**: Test-retest reliability was not accomplished. Given the rapidity of the downhill course in this disease, repeat testing would be needed within two weeks. We are now planning

to gather those data. We have not tested the SBRS on normal or cognitively impaired children other than those with Sanfilippo syndrome. Test results on a group of cognitive impaired children have been gathered in the UK, and we are hoping to be able to obtain those data for comparison.

Lessons Learned

1. We learned that there is a unique behavioral phenotype for Sanfilippo syndrome and we found support for the identification of the neuropathology that underlies this behavioral abnormality. We are now finding that there are other conditions with unique behavioral phenotypes. For example, Gaucher Disease Type III in Egypt appears to have a phenotype that includes severe aggression, oppositional behavior, and conduct problems. We are working with colleagues in Egypt to develop a strategy for creating a behavioral measure. Other diseases may also lend themselves to this approach.
2. Creating a disease-specific scale is a time- and effort-intensive endeavor; financial support is necessary to do it right. Moreover, as scale development is an iterative process, continued work is needed to improve its sensitivity to behavioral change.
3. Utilization of parental input and patient advocacy groups was crucial to our success. Parents were extraordinarily helpful and the cooperation of the MPS Society was crucial to carry out this work. During the time that parents were completing the measures, many asked us questions, made suggestions, and discussed their children's behavior with us. These interactions greatly improved the quality of the scale.
4. A team approach is necessary to implement such a study. The team must include parents, patient advocacy groups, patients (whenever possible), behavioral experts, statisticians, and (in our case) imaging specialists.

Tools to Standardize Assessments Across Multi-Site Trials

Methods to Improve Standardization of Neuropsychological Assessment in Clinical Trials

Kathleen A. Delaney

Standardized, validated measures are essential for evaluating outcomes in clinical trials. This is especially true for clinical trials requiring neurocognitive assessments. In addition to neurocognitive assessments, evaluation of behavior and ability to function on a day-to day basis may also be included. Such measures, however, can be vulnerable to a variety of factors that can increase the variability in outcome. These factors are listed below with solutions to decrease variability.

1. Choosing the correct assessment measure is critical for reducing variability. Limited patient populations that are widely dispersed geographically make clinical trials in IEMs particularly challenging. Trials in pediatric IEMs sometimes need to be global to meet enrollment goals for patient recruitment. Variability in assessment experience, approaches to testing, and cultural-specific test factors across sites can be problematic for gathering standardized data and for analyzing test results.

Using tools known and validated for global use results in greater reliability across sites and improved data quality. The test needs to be appropriate for the age group, developmental level, and disease in question. In choosing tests for a disease, population presence of severe cognitive impairment, sensory, motor, or behavior problems should be considered. For children who are severely impaired, often tests need to be used without standardized scores but can yield age-equivalent scores based on normative data.[5] For older or less impaired children, if the trial is international, nonverbal scales yield more reliable data. Use of language-based tests requires language-specific normative data. For each disease, in order to develop a set of common data elements, protocols with easily administered, but validated measures should be used

across trials. Using tests with up-to-date normative data is important. Natural history study data is critical for test selection as these data can help to determine the sensitivity of a test to change in a specific disease population.

The test protocol needs to be short in order to decrease fatigue that can also contribute to variable results. The time allotted to testing should vary with the age of the child but even in older children should not exceed two or three hours. Making sure that the child is using prescribed glasses and hearing aids, has had adequate sleep prior to testing, and is in an appropriate environment for testing ensures less variable results.

Practice effects are a concern in clinical trials that repeat testing frequently. Test repetition is not a significant problem in children under 4 because their rapid developmental course requires that different items be administered at each testing or their impairment precludes learning over time. In older children, cognitive testing should not be repeated more frequently than every 6 months. For memory testing in older children, measures with alternate forms are recommended. Some measures such as continuous performance measures of attention do not have significant practice effects.

2. Training of testers will significantly reduce variability. Testers who have experience with the patient populations are critically important; experience with administering similar measures is also important. Training across sites and assessors reduces variability in data and increases data quality. Allowing for some flexibility but maintaining test validity is key and is a skill achieved only by the most experienced, highly trained assessors. Enlisting the services of trained professionals with expertise and experience in the disease being studied is essential for achieving the most reliable data and accurate study outcomes. For a multisite clinical trial, the same person must train all testers making sure first that the tester has the fundamentals of testing in children, healthy or not. Then the trainer ensures that the tester becomes very familiar with the test. The trainer instructs the tester to know the test well and practice on control subjects in order to move smoothly through the test administration. The trainer provides examples of how to handle specific situations while staying within the guidelines of the test standardization. Instruction regarding behavioral modification techniques, rewards, when to involve parents, when to terminate the testing, and decision making about validity of the data collected are important topics. Video can be used to ensure consistency across sites.

3. Quality control will ensure accuracy and decrease variability. For scoring accuracy, test scoring software or centralized scoring can minimize the risk for error. Clinical trial monitors need to become familiar with the tools in order to identify issues of accuracy. For the testers, video may be used to review assessment accuracy. Alternatively, the trainer can visit sites to discuss issues with testers on a regular basis.

In conclusion, systematically reducing variability around assessment issues will improve the data collection process and the reliability and usefulness of trial results. Planning ahead with appropriate test selection and language considerations is critical. Researchers studying Inborn Errors of Metabolism must recognize the need for validated, disease-specific outcome measures. Training across sites with experienced raters can reduce the amount of variability in neurocognitive data quality. Complementary to disease-specific outcome measures are common data elements (discussed later in this article) that may be useful in studying a variety of rare diseases.

Common Data Element Project of the NIH National Institute of Neurological Disorders and Stroke

Joanne Odenkirchen, MPH, Clinical Research Project Manager,
Office of Clinical Research, Office of the Director
National Institute of Neurological Disorders and Stroke (NINDS), National Institutes of Health

Description, Purpose, and Scope of the NINDS Common Data Elements Project

The ability to accurately assess the effects of treatment for rare diseases requires reliable, clinically validated measures. This is especially true for inborn errors of metabolism (IEMs), where clinical researchers must assess neurocognitive abilities. The wide variety (in terms of etiology and phenotype) of IEMs and the small numbers of patients make research challenging. Common data elements (CDEs) that enable researchers studying different IEMs to readily share information could increase understanding of these diseases and ultimately expand treatment options.

The NINDS Common Data Elements Project enables clinical investigators to systematically collect, analyze, and share data across research communities. This ongoing project uses content standards for data elements that apply to multiple fields of research. For new projects, NINDS identifies disease-specific experts worldwide to collaborate to identify CDEs and instruments that are relevant to clinical research in that disease. Working groups work independently for 6-9 months, with NINDS providing technical advice and guidance upon request. Once CDEs are recommended by working groups, they are put out for public review and then posted on the NINDS CDE Web site for all to use. An Oversight Committee regularly reviews the existing CDEs, as the elements are dynamic. The Committee provides recommendations to NINDS on revisions/updates to the CDEs based on new scientific input. The NINDS CDE project currently has over 10,000 CDEs and 550 instruments in 19 diseases related to neurology. The NINDS project collaborates with other NIH, Federal, non-profit, and international data standard activities.

The NINDS CDE Project identifies common definitions and standardizes case report forms and other instruments to help investigators conduct clinical research using uniform formats. This project enables researchers to systematically collect, analyze and share clinical data across research communities.

Objectives of the NINDS Common Data Elements Project

The NINDS Common Data Elements Project aims to establish a transparent and inclusive process for data collection and sharing that can inform and connect multiple fields of research. For studies of rare diseases including Inborn Errors of Metabolism, the cross-pollination provided by CDEs could advance research opportunities in a challenging field for which limited data has been available. The ultimate goals include better documenting natural history of diseases and more accurately measuring treatment outcomes in order to provide the most effective treatments to support patients and their families.

The objectives of the CDE Project are to:

a. Identify CDEs used in clinical research
b. Present data elements in standard formats available to all
c. Identify common definitions and, where possible, validate permissible values and ranges that can guide researchers in selecting CDEs most applicable to their studies.
d. Standardize case report forms and other instruments
e. Provide standardized information to researchers for database development
f. Draw on expertise within the scientific community to identify, develop, and vet CDEs
g. Offer continuous support and guidance from NINDS and NINDS CDE Team
h. Maintain an iterative process that includes plans to annually review and update CDEs
i. Form an Oversight Committee (OC) to help maintain disease-specific CDEs

Common data elements are dynamic and will continue to evolve over time. New investigators can build on common data elements that are developed by consensus from members of the scientific community, facilitating multi-center and international clinical research efforts. The NINDS vision for the future of the CDE project includes:

a. Future NINDS-funded trials use CDEs or be CDE-compatible. This is already part of NINDS funding announcements and Terms of Award. Other NIH Institutes and Centers encourage CDEs where appropriate
b. All types of clinical research can use parts of the CDEs
c. Observational clinical studies can be linked to trial datasets
d. All NINDS human subject grantees are asked to consider using CDEs
e. Clinical research progress will be accelerated
f. NINDS is collaborating with the NLM to develop a CDE Hybrid Model through the NLM CDE Repository which hosts all of the NIH CDE; it is a platform for identifying related data elements in use across diverse areas, for harmonizing data elements, and for linking CDEs to other existing standards and terminologies.

NINDS expects the clinical research it funds to meet the highest standards of scientific rigor yet appreciates the burden that extensive data collection puts on investigators and study participants. Further, the Institute leadership recognizes that investigators independently identify data elements and forms for each study, many of which could be common across studies. As part of its effort to facilitate research of the highest quality yet streamline clinical trial data collection in neurological studies, in 2007 NINDS initiated the Common Data Element Project. NINDS will continue to expand on version 1.0 and refine the process for establishing and validating CDEs with the goal of accelerating progress in clinical research on a variety of diseases, including rare diseases such as Inborn Errors of Metabolism.

Common data elements are important tools for furthering research. This is especially true for ultra-rare diseases in which different researchers are collecting data on similar patient groups, often with different measures, rendering the data less comparable. It is critical for investigators to combine data across research centers using agreed-upon common data elements in order to collect sufficient natural history data that can serve as reference points in treatment trials.

Using Remote Technology to Expand the Reach of Clinical Research

Heather R. Adams, PhD., Associate Professor of Neurology and Pediatrics,
University of Rochester Medical Center

Batten Disease and the Challenges of Clinical Research

The neuronal ceroid lipofuscinoses (NCLs) are neurodegenerative, lysosomal storage diseases with predominantly childhood onset.[45] Despite distinct genetic etiologies and variable age of onset, most NCLs have similar clinical phenotypes that typically involve vision loss, motor decline, seizures, and dementia including cognitive decline and mood and behavioral symptoms.[46] [Santavuori] Juvenile neuronal ceroid lipofuscinosis (*CLN3*, also termed, 'juvenile Batten disease') is autosomal recessively inherited, affects males and females equally, and usually has onset between 4-7 years of age; children typically survive until their third decade.[47] [Mink, Augustine et al 2013] With an estimated worldwide incidence of 0.2 to 7.0 per 100,000 depending on geographic region and mutation pattern, juvenile Batten disease is considered a rare disease.[48, 49] Longer disease duration is associated with declines in physical, adaptive, and cognitive function, and patient and family-level quality of life is associated with behavioral and mood symptoms.[29, 30, 50, 51]

Because juvenile Batten disease is rare and medically complex, clinical research is challenging. There are relatively few affected individuals and likewise few clinical researchers with disease expertise, and these two groups (patients and researchers) are geographically disparate from one another. Participation in research may be further restricted because of the travel burden for children who are blind, have behavioral and cognitive symptoms of dementia, experience seizures, and eventually lose ambulation. The University of Rochester Batten Center (URBC) has therefore utilized a mixed approach of remote and direct assessments to evaluate clinical features and natural history of juvenile Batten disease, including assessment of behavior and cognition.

What Is Remote Assessment?

Remote assessment may involve any assessment in which the evaluator and examinee are not located within the same physical space. Examples include mail-in surveys, phone-based assessments directed by an examiner or administered automatically with touch-tone style response options, evaluations conducted through videoconferencing, or computer-generated testing that examinees access remotely through an internet-based connection and without any synchronous 'live' interaction with an examiner.[52-60] For videoconferencing, subjects may remain in their home or travel to a local physician's office or other designated site, where they can 'dial in' with the study team. In fact, some forms of behavioral assessment (e.g., standardized parent or teacher proxy reports of a subjects' behavior) are well-suited for remote assessment, as they are already designed to be completed independently by the respondent and are only returned to the assessor (via mail or at the end of an in-person visit) once completed.[61] Some of these measures have already been technologically updated so they may be completed electronically rather than on paper.[62] These and cognitive evaluations via remote methods have been utilized in both adult and pediatric studies, and have been demonstrated to be reliable and valid, with remotely obtained data comparable to those obtained during direct (face-to-face) assessments.

The URBC has also collected behavioral and adaptive function information about juvenile Batten disease via survey and phone interview for over 10 years[30], which has permitted repeated (every 6 months) assessment of children who would be unable to travel to our research frequently due to disease and financial burden. Additionally, we have demonstrated feasibility and reliability of live, remote video assessment for completion of the Unified Batten Disease Rating Scale (UBDRS), a disease-specific physical exam for Batten disease, and a recent pilot study demonstrated feasibility and high inter-rater agreement for remote neurocognitive assessment of affected individuals.[34, 63]

Advantages of Remote Assessment

Remote assessment may afford several advantages for clinical research activities, including improving time- and cost effectiveness and enhancing participant satisfaction.[54, 64] Remote assessment can significantly reduce travel burden, and the experience of families who travel to the URBC for each research visit was a motivating factor for us in developing an approach to 'virtual visits' for research subjects. To illustrate, the average round-trip travel distance for a family attending a single research visit at the URBC (from 2005-2014) was over 1,700 miles (SD approximately 1,300). These visits typically included overnight stays in Rochester, and parents may have had to arrange their own work schedules, care for other children, and transportation support for a disabled child, in order to visit the URBC for a half-day clinical research visit.

Data quality is a particular concern in evaluating individuals for clinical trials. This concern may be amplified in rare disease clinical research, where there may be very few qualified examiners with both disease and assessment expertise and experience.[65, 66] Centralized assessment by a limited number of trained raters is a potential solution, which may be supported by remote technology. As well, by reducing patient burden (through minimizing or eliminating travel and its attendant physical demands), subjects' attention, effort, and motivation during cognitive testing may be optimized.

Beyond the potential to evaluate cognition and behavior, remote methods may have broader applicability to other components of the clinical research visit. In many instances, remote approaches may already be in use, such as telephone calls to obtain safety and interim data, gather patient-reported outcomes, or even conduct initial recruitment campaigns (e.g., ResearchMatch.org). A recent study of spinal muscular atrophy, another IEM, illustrated the feasibility of 'virtual visits' for clinical research in rare diseases.[67] There is also emerging work on the use of remote delivery (i.e., 'telemedicine') of psychosocial and behavioral clinical trial interventions,[68, 69] and to increase patient access to specialist care.[70, 71] It remains to be seen whether remote methods, including components of interventions, can be extended more broadly to clinical research in the IEMs and other rare diseases, while always maintaining the highest emphasis on patient safety. In addition, we recognize that sensory and motor impairments, developmental state (e.g., very young children), and the goal of the assessment may preclude the use of a 'virtual visit' in some instances. For assessments of cognition and behavior in particular, psychometric standards may need to be re-established when test administration changes to remote-based methods. Maintaining a controlled study environment is critical, particularly for standardized cognitive evaluations. Additionally, there are technological barriers in some cases (absent or insufficient Internet service, hardware, or software), privacy concerns related to electronic transmission of information, and cultural and generational considerations that will need to be identified and addressed. With an estimated 7,000 rare diseases affecting up to 30 million individuals around the globe, access to clinical research (including trials) and to disease experts is significant. The URBC is currently implementing a study of patient and caregiver-reported barriers to clinical research, and whether remote technology may aid in removing such barriers. This study expands beyond Batten disease to encompass numerous rare health conditions.

Conclusion

There are myriad reasons why clinical studies of Batten disease and other inborn errors of metabolism are particularly challenging. Remote technologies offer the opportunity to address some of these challenges. Patients with rare diseases are typically in disparate locations, making travel to centralized research centers costly, time-consuming, and difficult. Remote assessment offers the proficiency of researchers who are experts in the disease and in assessment measures, and in settings that are convenient and comfortable for patients and their families. By reducing the number of assessors, increasing measurement expertise, and providing a more controlled environment, remote assessment may improve data quality and rigor and reduce sources of variability. In the long term, the ability to use remote technologies may increase patient engagement in research and foster new discoveries and advances in patient care. Overall, we view remote assessment as an additional tool in the repertoire of clinical researchers, to be used as and when appropriate to engage and retain participants in various clinical research activities, enhance access to clinical research, and ultimately improve access, timeliness, and quality of care for rare diseases.

References

[1] D.N. Gliklich R, Leavy M, eds. , Registries for Evaluating Patient Outcomes: A User's Guide. Third edition. Two volumes. (Prepared by the Outcome DEcIDE Center [Outcome Sciences, Inc., a Quintiles company] under Contract No. 290 2005 00351 TO7.) AHRQ Publication No. 13(14)-EHC111. Rockville, MD: Agency for Healthcare Research and Quality. (April 2014).

[2] S.C. Groft, M.P. de la Paz, Rare diseases - avoiding misperceptions and establishing realities: the need for reliable epidemiological data Adv Exp Med Biol 686 (2010) 3-14.

[3] Rare Diseases: Common Issues in Drug Development -- Guidance for Industry, Food and Drug Administration, August 2015.

[4] C. Peters, E.G. Shapiro, J. Anderson, P.J. Henslee-Downey, M.R. Klemperer, M.J. Cowan, E.F. Saunders, P.A. deAlarcon, C. Twist, J.B. Nachman, G.A. Hale, R.E. Harris, M.K. Rozans, J. Kurtzberg, G.H. Grayson, T.E. Williams, C. Lenarsky, J.E. Wagner, W. Krivit, Hurler syndrome: II. Outcome of HLA-genotypically identical sibling and HLA-haploidentical related donor bone marrow transplantation in fifty-four children. The Storage Disease Collaborative Study Group Blood 91 (1998) 2601-2608.

[5] K. Delaney, K. Rudser, B. Yund, C. Whitley, P.J. Haslett, E. Shapiro, Methods of Neurodevelopmental Assessment in Children with Neurodegenerative Disease: Sanfilippo Syndrome, in: J. Zschocke, K.M. Gibson, G. Brown, E. Morava, V. Peters (Eds.), JIMD Reports - Case and Research Reports, Volume 13, Springer Berlin Heidelberg, 2014, pp. 129-137.

[6] D.J. Raggio, T.W. Massingale, J.D. Bass, Comparison of Vineland Adaptive-Behavior Scales Survey Form Age Equivalent and Standard Score with the Bayley Mental-Development Index Percept Motor Skill 79 (1994) 203-206.

[7] N. Bayley, Bayley scales of infant and toddler development– third edition. Psychological Corporation, San Antonio (2006).

[8] S.S. Sparrow, Chicchetti, D.V.; Balla, D.A., Vineland Adaptive Behavior Scales, 2nd edition Psychological Corporation, San Antonio TX (2005).

[9] E. Shapiro, O.E. Guler, K. Rudser, K. Delaney, K. Bjoraker, C. Whitley, J. Tolar, P. Orchard, J. Provenzale, K.M. Thomas, An exploratory study of brain function and structure in mucopolysaccharidosis type I: long term observations following hematopoietic cell transplantation (HCT) Mol Genet Metab 107 (2012) 116-121.

[10] E.G. Shapiro, I. Nestrasil, K.A. Delaney, K. Rudser, V. Kovac, N. Nair, C.W. Richard, 3rd, P. Haslett, C.B. Whitley, A Prospective Natural History Study of Mucopolysaccharidosis Type IIIA J Pediatr 170 (2016) 278-287 e274.

[11] P. Harrison, T. Oakland, Adaptive Behavior Assessment System, 2nd Edition The Psychological Corporation, San Antonio TX (2003).

[12] S.E. Waisbren, J. He, R. McCarter, Assessing Psychological Functioning in Metabolic Disorders: Validation of the Adaptive Behavior Assessment System, Second Edition (ABAS-II), and the Behavior Rating Inventory of Executive Function (BRIEF) for Identification of Individuals at Risk JIMD Rep 21 (2015) 35-43.

[13] B.M. Craig, B.B. Reeve, P.M. Brown, D. Cella, R.D. Hays, J. Lipscomb, A. Simon Pickard, D.A. Revicki, US valuation of health outcomes measured using the PROMIS-29 Value Health 17 (2014) 846-853.

[14] D. Wechsler, Wechsler Abbreviated Scale of Intelligence, 2nd Edition. The Psychological Corporation, San Antonio TX (2011).

[15] D. Wechsler, Wechsler Preschool and Primary Scale of Intelligence. 4th Edition The Psychological Corporation, San Antonio TX (2012).

[16] R.L. Achenbach TM, Manual for the ASEBA School-Aged Froms and Profiles University of Vermont, Research Center for Children, Youth, and Families, Burlington VT (2001).

[17] K.E. Beery, Buktenica, N.A.; Beery, N.A. , Beery-Buktenica Developmental Test of Visual-Motor Integration (VMI) - 6th Edition. Pro-Ed, Austin TX. (2004).

[18] D.C. Delis, K. J.H., E. Kaplan, B.A. Ober, California Verbal Learning Test, Children's Version. The Psychological Corporation (1994).
[19] D.C. Delis, K. J.H., E. Kaplan, B.A. Ober, California Verbal Learning Test, 2nd Edition The Psychological Corporation (2000).
[20] L.I. Company, Grooved Pegboard Model 32025 Lafayette IN (1989).
[21] J. Muenzer, Early initiation of enzyme replacement therapy for the mucopolysaccharidoses Mol Genet Metab 111 (2014) 63-72.
[22] E.G. Shapiro, I. Nestrasil, A. Ahmed, A. Wey, K.R. Rudser, K.A. Delaney, R.K. Rumsey, P.A. Haslett, C.B. Whitley, M. Potegal, Quantifying behaviors of children with Sanfilippo syndrome: the Sanfilippo Behavior Rating Scale Mol Genet Metab 114 (2015) 594-598.
[23] D.C. De Vivo, R.R. Trifiletti, R.I. Jacobson, G.M. Ronen, R.A. Behmand, S.I. Harik, Defective glucose transport across the blood-brain barrier as a cause of persistent hypoglycorrhachia, seizures, and developmental delay New England Journal of Medicine 325 (1991) 703-709.
[24] T.S. Pearson, C. Akman, V.J. Hinton, K. Engelstad, C. Darryl, Phenotypic spectrum of glucose transporter type 1 deficiency syndrome (Glut1 DS) Current neurology and neuroscience reports 13 (2013) 1-9.
[25] A.W. Pong, B.R. Geary, K.M. Engelstad, A. Natarajan, H. Yang, D.C. De Vivo, Glucose transporter type I deficiency syndrome: epilepsy phenotypes and outcomes Epilepsia 53 (2012) 1503-1510.
[26] R. Pons, A. Collins, M. Rotstein, K. Engelstad, D.C. De Vivo, The spectrum of movement disorders in Glut-1 deficiency Mov Disord 25 (2010) 275-281.
[27] F.J. Marshall, E.A. de Blieck, J.W. Mink, L. Dure, H. Adams, S. Messing, P.G. Rothberg, E. Levy, T. McDonough, J. DeYoung, M. Wang, D. Ramirez-Montealegre, J.M. Kwon, D.A. Pearce, A clinical rating scale for Batten disease: reliable and relevant for clinical trials Neurology 65 (2005) 275-279.
[28] E.A. de Blieck, E.F. Augustine, F.J. Marshall, H. Adams, J. Cialone, L. Dure, J.M. Kwon, N. Newhouse, K. Rose, P.G. Rothberg, A. Vierhile, J.W. Mink, Methodology of clinical research in rare diseases: development of a research program in juvenile neuronal ceroid lipofuscinosis (JNCL) via creation of a patient registry and collaboration with patient advocates Contemp Clin Trials 35 (2013) 48-54.
[29] H. Adams, C. Beck, E. Levy, R. Jordan, J. Kwon, F. Marshall, A. Vierhile, E. Augustine, E. deBlieck, D. Pearce, M. JW, Genotype does not predict severity of behavioral phenotype in Juvenile Neuronal Ceroid Lipofuscinosis (Batten Disease) Developmental Medicine & Child Neurology 52 (2010) 637-643. PMID:20187884.
[30] H. Adams, J. Mink, and the University of Rochester Batten Center Study Group, Neurobehavioral Features and Natural History of Juvenile Neuronal Ceroid Lipofuscinosis (Batten Disease) Journal of child neurology 28 (2013) 1128-1136.
[31] H.R. Adams, J. Kwon, F.J. Marshall, E.A. de Blieck, D.A. Pearce, J.W. Mink, Neuropsychological symptoms of juvenile-onset batten disease: experiences from 2 studies Journal of child neurology 22 (2007) 621-627.
[32] J.M. Kwon, H. Adams, P.G. Rothberg, E.F. Augustine, F.J. Marshall, E.A. Deblieck, A. Vierhile, C.A. Beck, N.J. Newhouse, J. Cialone, E. Levy, D. Ramirez-Montealegre, L.S. Dure, K.R. Rose, J.W. Mink, Quantifying physical decline in juvenile neuronal ceroid lipofuscinosis (Batten disease) Neurology 77 (2011) 1801-1807.
[33] E.F. Augustine, H.R. Adams, C.A. Beck, A. Vierhile, J. Kwon, P.G. Rothberg, F. Marshall, R. Block, J. Dolan, J.W. Mink, Standardized assessment of seizures in patients with juvenile neuronal ceroid lipofuscinosis Dev Med Child Neurol 57 (2015) 366-371.
[34] Cialone J, Augustine EF, Newhouse N, Vierhile A, Marshall FJ, M. JW, Quantitative telemedicine ratings in Batten disease: implications for rare disease research Neurology 77 (2011) 1801-1811.

[35] J. Cialone, H. Adams, E.F. Augustine, F.J. Marshall, J.M. Kwon, N. Newhouse, A. Vierhile, E. Levy, L.S. Dure, K.R. Rose, D. Ramirez-Montealegre, E.A. de Blieck, J.W. Mink, Females experience a more severe disease course in Batten disease J Inherit Metab Dis 35 (2012) 549-555.
[36] J. Cialone, E.F. Augustine, N. Newhouse, H. Adams, A. Vierhile, F.J. Marshall, E.A. de Blieck, J. Kwon, P.G. Rothberg, J.W. Mink, Parent-reported benefits of flupirtine in juvenile neuronal ceroid lipofuscinosis (Batten disease; CLN3) are not supported by quantitative data J Inherit Metab Dis 34 (2011) 1075-1081.
[37] C. Nguyen, E.R. Foster, A.R. Paciorkowski, A. Viehoever, C. Considine, A. Bondurant, B.A. Marshall, T. Hershey, Reliability and validity of the Wolfram Unified Rating Scale (WURS) Orphanet J Rare Dis 7 (2012) 89.
[38] M.A. Cleary, J.E. Wraith, Management of mucopolysaccharidosis type III Arch Dis Child 69 (1993) 403-406.
[39] G.A. Colville, M.A. Bax, Early presentation in the mucopolysaccharide disorders Child Care Health Dev 22 (1996) 31-36.
[40] T.C. Neylan, Temporal lobe and behavior: Kluver and Bucy's classic J Neuropsych Clin N 9 (1997) 606-606.
[41] K.A. Buss, H.H. Goldsmith, Manual and normative data for the Laboratory Temperament Assessment Battery—Toddler Version Tech. Rep, Department of Psychology, University of Wisconsin, Madison, WI (2000).
[42] K. Gotham, S. Risi, A. Pickles, C. Lord, The Autism Diagnostic Observation Schedule: revised algorithms for improved diagnostic validity J Autism Dev Disord 37 (2007) 613-627.
[43] M. Potegal, B. Yund, K. Rudser, A. Ahmed, K. Delaney, I. Nestrasil, C.B. Whitley, E.G. Shapiro, Mucopolysaccharidosis Type IIIA presents as a variant of Kluver-Bucy syndrome J Clin Exp Neuropsychol 35 (2013) 608-616.
[44] R.K. Rumsey, K. Rudser, K. Delaney, M. Potegal, C.B. Whitley, E. Shapiro, Acquired autistic behaviors in children with mucopolysaccharidosis type IIIA J Pediatr 164 (2014) 1147-1151 e1141.
[45] R. Williams, S. Mole, New nomenclature and classification scheme for the neuronal ceroid lipofuscinoses Neurology 79 (2012) 183-191.
[46] P. Santavuori, L. Lauronen, E. Kirveskari, L. Åberg, K. Sainio, T. Autti, Neuronal ceroid lipofuscinoses in childhood Neurol. Sci. 21 (2000) S35-41.
[47] J.W. Mink, E.F. Augustine, H.R. Adams, F.J. Marshall, J.M. Kwon, Classification and natural history of the neuronal ceroid lipofuscinoses Journal of child neurology 28 (2013) 1101-1105.
[48] R. Williams, NCL incidence and prevalence data, in: S. Mole, R. Williams, H. Goebel (Eds.), The neuronal ceroid lipofuscinoses: Batten Disease, Oxford University Press, Oxford, UK, 2011.
[49] "Rare Diseases Act of 2002" (PL 107-280, Nov 6, 2002). 116 Stat 1988, 2002.
[50] H. Adams, E.A. de Blieck, J.W. Mink, F.J. Marshall, J. Kwon, L. Dure, P.G. Rothberg, D. Ramirez-Montealegre, D.A. Pearce, Standardized assessment of behavior and adaptive living skills in juvenile neuronal ceroid lipofuscinosis Dev Med Child Neurol 48 (2006) 259-264.
[51] H. Adams, E. deBlieck, J. Kwon, F. Marshall, D. Pearce, P. Rothberg, J. Mink, Quality of life in Batten Disease: Affected children, parents, and siblings, 11th International Congress on Neuronal Ceroid Lipofuscinosis (Batten Disease), Rochester, NY, July 2007.
[52] T. Castanho, L. Amorim, J. Zihl, J. Palha, N. Sousa, N. Santos, Telephone-based screening tools for mild cognitive impairment and dementia in aging studies: a review of validated instruments Front Aging Neurosci. 25 (Epub 25 Feb 2014).
[53] P. Dale, N. Harlaar, R. Plomin, Telephone testing and teacher assessment of reading skills in 7-year olds: substantial correspondence for a sample of 5544 children and for extremes Read Writ 18 (2005) 385-400.
[54] E.R. Dorsey, C. Venuto, V. Venkataraman, D.A. Harris, K. Kieburtz, Novel methods and technologies for 21st-century clinical trials: a review JAMA neurology 72 (2015) 582-588.
[55] J. Kent, R. Plomin, Testing specific cognitive abilities by telephone and mail Intelligence 11 (1987) 391-400.

[56] S. Petrill, J. Rempell, B. Oliver, R. Plomin, Testing cognitive abilities by telephone in a sample of 6- to 8-year olds Intelligence 30 (2002).
[57] V. Temple, C. Drummond, S. Valiquette, E. Jozsvai, A comparison of intellectual assessments over video conferencing and in-person for individuals with ID: preliminary data J Intell Disabil Res 54 (2010) 573-577.
[58] M. Waite, D. Theodoros, T. Russell, L. Cahill, Assessment of children's literacy via an internet-based telehealth system Telemed J E-Health 16 (2010) 564-575.
[59] M. Waite, D. Theodoros, T. Russell, L. Cahill, Internet-based telehealth assessment of language using the CELF-4 Lang Speech Hear Ser 41 (2010) 445-458.
[60] S. Walsh, R. Raman, K. Jones, P. Aisen, for the Alzheimer's Disease Cooperative Study Group, ADCS prevention instrument project: The Mail-In Cognitive Function Screening Instrument (MCFSI) Alz Dis Assoc Dis 20 (2006) S170-178.
[61] S. Whitcomb, K. Merrell, Behavior rating scales, in: S. Whitcomb, K. Merrell (Eds.), Behavioral, Social, and Emotional Assessment of Children and Adolescents, Routledge, New York, NY, 2013, pp. 127-158.
[62] D. Richard, D. Lauterbach, Computers in the training and practice of behavioral assessment, in: M. Hersen (Ed.), Comprehensive Handbook of Psychological Assessment. Volume 3: Behavioral Assessment. , 2004, pp. 222-245.
[63] S.N. Ragbeer, E.F. Augustine, J.W. Mink, A.R. Thatcher, A.E. Vierhile, H.R. Adams, Remote assessment of cognitive function in juvenile neuronal ceroid lipofuscinosis (Batten disease) – a pilot study of feasibility and reliability Journal of child neurology in press (2015).
[64] E. Dorsey, V. Venkataraman, M. Grana, M. Bull, B. George, C. Boyd, C. Beck, B. Rajan, A. Seidmann, K. Biglan, Randomized controlled clinical trial of "virtual house calls" for Parkinson disease JAMA neurology 70 (2013) 565-570.
[65] K. Kobak, N. Engelhardt, W. JBW, J. Lipsitz, Rater Training in Multicenter Clinical Trials: Issues and Recommendations J Clin Pharmacol 24 (2004) 113-117.
[66] K. Kobak, J. Kane, M. Thase, A. Nierenberg, Why Do Clinical Trials Fail? The Problem of Measurement Error in Clinical Trials: Time to Test New Paradigms? J Clin Pharmacol 27 (2007) 1-5.
[67] T. Chen, Y. Yang, H. Mai, W. Liang, Y. Wu, H. Wang, Y. Jong, Reliability and validity of outcome measures of in-hospital and at-home visits in a randomized, double-blind, placebo-controlled trial for spinal muscular atrophy Journal of child neurology 29 (2014) 1680-1684.
[68] N.C. Slone, R.J. Reese, M.J. McClellan, Telepsychology outcome research with children and adolescents: a review of the literature Psychological services 9 (2012) 272-292.
[69] D.M. Hilty, D.C. Ferrer, M.B. Parish, B. Johnston, E.J. Callahan, P.M. Yellowlees, The effectiveness of telemental health: a 2013 review Telemedicine journal and e-health : the official journal of the American Telemedicine Association 19 (2013) 444-454.
[70] M. Achey, C. Beck, D. Beran, C. Boyd, P. Schmidt, A. Willis, S. Riggare, R. Simone, K. Biglan, E. Dorsey, Virtual house calls for Parkinson disease (Connect.Parkinson): study protocol for a randomized controlled trial Trials (Epub date 27 Nov 2014).
[71] E. Dorsey, L. Deuel, T. Voss, K. Finnigan, B. George, S. Eason, D. Miller, J. Reminick, A. Appler, J. Polanowicz, L. Viti, S. Smith, A. Joseph, K. Biglan, Increasing access to specialty care: a pilot randomized controlled trial of telemedicine for Parkinson's disease J Telemed Telecare 15 (2010) 115-117.