Clinical Outcome Assessments: Use of Normative Data in a Pediatric Rare Disease

Dawn Phillips, BScPT, MS, PhD1,2,*; Beth Leiro, BScPT1

1Evidera Inc., Bethesda, MD, USA; 2UNC Division of Physical Therapy, Chapel Hill, NC, USA

ABSTRACT

Pediatric rare diseases present unique challenges in clinical trial design and in selection of clinical outcome assessments (COAs) used to support claims in medical product labeling. COAs that discriminate level of function relative to a normative sample are particularly important in the pediatric rare disease setting because the literature is often void of natural history data. Pediatric rare disease clinical trials will often include a wide age distribution. Gross and fine motor skills, communication, cognition, and independence in activities of daily living vary by age, and it may be difficult to distinguish between treatment effect and change due to developmental maturation. Asfotase alfa was granted breakthrough therapy designation and subsequently approved for the treatment of hypophosphatasia (HPP; a genetic metabolic musculoskeletal disorder) and is used in this discussion to illustrate COA selection in a pediatric rare disease. Multiple COAs with normative data in HPP clinical trials for asfotase alfa are presented. The assessment instruments included the Bayley Scales of Infant and Toddler Development-Third Edition, the Bruininks-Oseretsky Test of Motor Proficiency, Second Edition, the Childhood Health Assessment Questionnaire, the Pediatric Outcomes Data Collection Instrument, handheld dynamometry, the 6-minute walk test, and the Modified Performance-Oriented Mobility Assessment-Gait scale. Multiple end points were required to adequately capture the impact of asfotase alfa treatment on the multiple systems affected in HPP. These data illustrate the importance of using multiple COAs that provide normative data and to use COAs early in the drug development process for rare pediatric disease.

Keywords: asfotase alfa, clinical outcomes, hypophosphatasia, rare disease.

Introduction

Clinical design in pediatric rare disease research often requires complicated end point models with multiple clinical outcome assessments (COAs) to capture the constructs. This commentary will focus on the use of existing standardized instruments that consider development by age and classify function relative to normative values. In this discussion, hypophosphatasia (HPP) is used to illustrate COA selection in a pediatric rare disease. HPP, a genetic metabolic musculoskeletal disorder, is caused by mutations in the tissue-nonspecific alkaline phosphatase gene [1]. Heterogeneous manifestations can include rickets, fractures, muscle weakness, limb deformities, pain, and respiratory compromise, which result in delayed acquisition of age-appropriate developmental skills, gait impairments, and decreased functional independence in activities of daily living (ADL) [1]. The disease has a particularly high burden in children and is associated with high mortality rates in infants [2].

Use of Existing Standardized Developmental Assessments

Rare disease studies often have a small sample size that is insufficient to divide by group-level differences in age and function. Normative data can define function in a heterogeneous sample and be used to characterize the disease presentation in the natural history and treatment groups. Rare diseases often present with multisystem impairments that limit ability to capture the direct impact of disease-defining concepts. Comprehensive developmental COAs can be used to characterize the disease impact on age-appropriate markers across multiple...
domains, examine the relationship between domains, and guide selection or development of additional disease-specific measures.

Pediatric rare disease clinical trials often enroll a wide age distribution. Gross and fine motor skills, communication, cognition, and independence in ADL vary by age, and it may be difficult to distinguish between treatment effects and change due to developmental maturation. Identical function may be age-appropriate for a younger child and considered atypical or delayed in an older child. Variability in typical function by age requires that multiple age versions be developed for disease-specific validated patient-reported outcomes. Comparisons with a normative sample in an existing tool can provide age-appropriate developmental expectations.

Existing standardized developmental assessments can provide a range of values to measure treatment benefits including, but not limited to, standard scores (including scale scores), percentile rank, age-equivalent (AE) scores, and developmental quotients (DQs) [3]. Different values may be used to interpret treatment benefit in infantile and juvenile disease phenotypes. The infantile form of the rare disease may include progressive loss of developmental skills and high mortality, and treatment benefit may be defined by the acquisition of a developmental skill that exceeds function observed in the natural history study. The rate of skill acquisition in response to a treatment may be slower than in the normative sample and improvement may be reflected only as an increase in AE scores. Standard scores can be insensitive to change in low-functioning children because either the children fall below the test floor or the rate of change is slower than in typically developing children in the normative sample, and standard scores either plateau or decline [3]. In a progressive condition in which treatment is focused on arresting deterioration, a treatment response may be indicated only by stable AE values because a subsequent decline in standard scores will result with increased age. DQ (AE/chronological age \times 100) can be sensitive to age as a determinant of disease progression in children with low function and can be used to compare the rate of change between disease natural history and intervention groups [5].

**Application to HPP**

The US Food and Drug Administration approved asfotase alfa (Strensiq®, Alexion Pharmaceuticals Inc., Cheshire, CT, USA) in 2015 as the first approved drug for perinatal, infantile, and juvenile HPP, after development with orphan drug designation. The clinical development plan included multiple studies to evaluate safety, tolerability, pharmacokinetics, pharmacodynamics, and efficacy. The end point model included HPP-related rickets as a primary end point and a combination of secondary/exploratory variables to provide a comprehensive picture of function, disability, pain, and health-related quality of life.

**Bayley Scales of Infant and Toddler Development—Third Edition**

The five developmental domains of the Bayley Scales of Infant and Toddler Development—Third Edition (Bayley-III) (i.e., cognition, language, motor skills, social-emotional behavior, and adaptive behavior) were developed (normed and validated) for use in impaired and healthy children aged between 1 and 42 months [4]. The Bayley-III is regarded as the best practice tool, and is thus most widely used in clinical practice [5]. In an open-label, retrospective study, 11 patients were assessed using Bayley-III at baseline and at 24 and 48 weeks after initiation of asfotase alfa for treatment of HPP [1,2]. All patients had fine motor, gross motor, and cognitive delays at baseline, and 87.5% of patients for whom data were available showed improvements in...
Table 1 – Roadmap to patient-focused outcome measurement in clinical trials—application to pediatric rare disease research and HPP.

<table>
<thead>
<tr>
<th>Challenges and Possible Strategies</th>
<th>Understanding the disease or condition</th>
<th>Conceptualizing treatment benefit</th>
<th>Selecting/developing the outcome measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Common challenges in pediatric rare disease research</td>
<td>• Literature void of natural history data, especially related to function</td>
<td>• Concept of interest for infantile presentation is often survival with an open-label single-arm study design</td>
<td>• Existing standardized developmental instruments can provide a measure of the impact of multisystem impairment and classify function relative to normative values, but</td>
</tr>
<tr>
<td></td>
<td>• Heterogeneity in disease presentation by phenotypes with variable age and functional presentations</td>
<td>• Conceptualizing benefit by how a child feels and functions is complicated because typical developmental function varies by age and involves a complex interaction between cognitive, communication, and motor skills</td>
<td>• Require extensive training for administration often within an international site distribution</td>
</tr>
<tr>
<td></td>
<td>• A broad inclusion of disease phenotypes allows better characterization for which therapy may be feasible, but adds increased design and analysis complexity</td>
<td>• Consideration must not only be given to the concepts of interest but also to the interactions between the concepts. Cognition, communication, or attentional capacity may limit ability to measure primary treatment benefit</td>
<td>• Challenging to establish disease-specific validation and that conceptual framework is appropriate for study population and end point</td>
</tr>
<tr>
<td></td>
<td>• Multisystem impairments</td>
<td>• May be difficult to distinguish between treatment effect and change due to developmental maturation. Identical function may be age-appropriate for a younger child and considered atypical or delayed in an older child. In the juvenile form of a rare disease, function may exceed the infantile presentation but the children have multiple comorbidities and the impact of the impairments need to be measured by comparison with age-appropriate task execution, social and peer interaction, and function within the home, school, and community environment</td>
<td>• Instrument manuals do not include guidelines for accommodations for special populations, such as strategies to obtain reliable neurocognitive assessments in children who are low-functioning, have noncooperative behavior, or have physical disabilities [22]</td>
</tr>
<tr>
<td></td>
<td>• International site distribution with variable standards of care</td>
<td>• Difficult to develop responder definition with heterogeneity in age and function</td>
<td>• Longitudinal data collection over years may require transitioning between developmental assessments with different psychometric properties</td>
</tr>
<tr>
<td></td>
<td>• Developing children have dynamic health states and impacts, so static measures are not sufficient to benchmark or assess the target population</td>
<td>• Open-label clinical trials in which patients and investigators are aware of assigned therapy are rarely adequate to support labeling claims based on PRO instruments alone. COAs that support improvement in specific symptoms would not support a general claim related to improvement, and multidomain claims cannot be substantiated by instruments that do not adequately measure the individual disease concepts. Clinical designs often require complicated end point models with multiple COA types to capture the constructs</td>
<td>• Standard scores can be used to discriminate function relative to SDs from the normative mean or percentile rank</td>
</tr>
<tr>
<td></td>
<td>• Infantile or severe disease presentations may include progressive loss of developmental skills and high mortality with no available treatment</td>
<td>• Targeted clinical trial populations are desirable for optimal design and ability to demonstrate treatment benefit, but are often limited by recruitment in rare disease and less desirable to have a narrow disease categorization for labeling</td>
<td>• May not show a treatment benefit (stable or increasing standard score) if new skills are acquired but at a slower rate than the normative sample. Age-equivalent scores may be more useful than standard scores to demonstrate skill acquisition in a child with severe motor impairment [22]</td>
</tr>
</tbody>
</table>

* Development of disease-specific validated PROs is challenging because of feasibility, time, and associated costs

* Motor skills and level of independence in ADL vary greatly by age and require validation of many items and multiple age versions

continued on next page
Potential strategies

- Use comprehensive prospective, observational, natural history studies with multiple COAs to gain insight into the multisystem impacts on age-appropriate markers (symptoms/impacts)
- Use these natural history studies to gain insights into the COA performance (sensitivity and specificity) and to look into the relationships between outcome measures (consider language, motor ability, and behavioral and cultural aspects)
- Characterize disease by distinct age and functional groups using natural history data, KOLs, patient and caregiver perspectives
- Treatment benefit in the infant population may be defined by global development and the pediatric/juvenile group may require performance-based or patient-reported assessments that are focused on a specific functional skill that is age-specific or disease-specific
- Treatment benefit may be defined by a responder definition on the basis of acquiring a developmental skill that exceeds function observed in the natural history study
- Treatment benefit may also be defined by DQs (age-equivalent scores/chronological age x 100) and compared with decline in DQ in the natural history study [22]
- Use KOL, focus group, caregiver, and patient perspectives to define treatment benefit
- Use assessment batteries with normative data for the age and culture being targeted
- Use disease-relevant domains of content within a developmental test
- Supplement the batteries with COA specific to the anticipated treatment benefits
- Develop standardized order for all COAs; evaluate areas of overlap between multiple performance instruments to reduce redundancy and subject fatigue
- Content validity—establishing evidence that an existing developmental instrument measures concepts of interests in rare disease
  - Highlight validation data used to develop instrument from diseases with similar impairments
  - Complete literature searches to support use in interventional studies with similar impairments
  - Use KOL perspective and consensus meetings to establish disease-specific recommendations for COA
  - Examine relationship between performance assessments and HRQOL in prospective observation study
- PRO instrument development—review the availability of item banks to derive items and to possibly expedite the development process. The PROMIS item banks may be a consideration because they have already had extensive field testing and are consistent with the International Classification of Function for Children and Youth [6]
- Develop responder definition on the basis of distribution analysis of groups in prospective observation study and expert, patient, and caregiver perspectives

HPP example

- Systematic literature searches and KOLs used to characterize distinct groups by age and function
- Retrospective natural history studies completed
  - Substudy of larger retrospective natural history study was conducted that assessed gait impairments from clinical gait videos
- Open-label, multinational, multicenter, single-arm design due to unmet medical need, serious mortality and morbidity risk, and absence of disease-modifying treatment
- Multiple studies to measure treatment benefit in infantile-, pediatric-, and adult-onset HPP
- End points used in infantile and pediatric studies Biochemical parameters
  - Tissue-nonspecific alkaline phosphatase enzyme substrates
  - Skeletal system measures
  - Bone mineralization-biopsy and DEXA

continued on next page
### Table 1 – continued

<table>
<thead>
<tr>
<th>Challenges and Possible Strategies</th>
<th>Understanding the disease or condition</th>
<th>Conceptualizing treatment benefit</th>
<th>Selecting/developing the outcome measure</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>• Multiple inter-related end points in each study that included PROs, ObsROs, ClinROs, and PerfOs</td>
<td>• Rickets severity</td>
</tr>
<tr>
<td></td>
<td></td>
<td>o The Rickets Severity Scale</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>o The Radiographic Impression of Change tool</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>o Growth Development function and strength—infantile</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>• BSID-III [9]</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Survival—respiratory status Physical function, strength, and ambulation—pediatric</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>• BOT-2: running speed and agility subtest and strength subtest</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>• HHD</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>• 6MWT</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>• MPOMA-G [6]</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Disability and HRQOL—pediatric</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>• CHAQ</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>• PODCI</td>
<td></td>
</tr>
</tbody>
</table>

6MWT, 6-minute walk test; ADL, activities of daily living; BOT-2, Bruininks-Oseretsky Test of Motor Proficiency-Second Edition; BSID-III, Bayley Scales of Infant Development-Third Edition; CHAQ, Childhood Health Assessment Questionnaire; ClinRO, clinician-reported outcomes; COA, clinical outcome assessments; DEXA, dual-energy x-ray absorptiometry; DQ, developmental quotient; HHD, handheld dynamometry; HPP, hypophosphatasia; HRQOL, health-related quality of life; KOL, key opinion leader; MPOMA-G, Modified Performance-Orientated Mobility Assessment-Gait; ObsRO, observer-reported outcomes; PerfO, performance outcomes; PODCI, Pediatric Outcomes Data Collection Instrument; PRO, patient-reported outcome; PROMIS, Patient-Reported Outcomes Measurement Information System.
these areas [1]. AE scores were used to illustrate linear skill acquisition, and scaled scores (mean 10 ± 3) illustrated rate and level of skill acquisition relative to a normative sample, with median (minimum, maximum) gross motor scaled scores increasing from 1 (1, 8) at baseline to 2 (1, 5) at week 48 [6].

**Bruininks-Oseretsky Test of Motor Proficiency-Second Edition**

The purpose of the Bruininks-Oseretsky Test of Motor Proficiency-Second Edition (BOT-2) is to assess motor skills, including differentiated measures of gross motor and fine motor proficiency, in individuals aged 4 to 21 years [7]. BOT-2 is a standardized sex-specific, norm-referenced, discriminative, and evaluative tool [7,8]. All items are administered to the entire age range and point scores are calculated from the raw scores for each item. Scaled scores and AE scores can be calculated from combined point scores for each subtest, and scaled scores can be combined to derive composite standard scores and percentile ranks [7].

It would not be appropriate to calculate mean point scores across a clinical trial with a wide age distribution because all the items within a subtest are administered to patients of all ages. Instead, scaled scores should be used because they are based on age-specific normative data [7,9]. For example, a running speed and agility point score of 35 for a 10-year-old child would equate to a scaled score of 15 (normative mean), and the same point score for a 17-year-old adolescent would equate to a scaled score of 10, which is 1 SD less than the normative mean for that age category.

In a phase II, open-label study of asfotase alfa in children aged 5 to 12 years (N = 13) with HPP, only the running speed and agility subtest and the strength subtest of BOT-2 were used because they were the most relevant to the disease-specific impairments and mobility restrictions and involved a reasonable amount of administrative time when paired with additional outcomes. At baseline, the median scaled score for both these BOT-2 subtests were more than 2 SDs below the normative mean. Asfotase alfa treatment resulted in significant and clinically meaningful improvements in strength (P < 0.0003) and running speed and agility (P < 0.0001), demonstrated by improvements in BOT-2 mean scores to +1 SD of normal at last assessment (week 144 or 168). The treatment had a proximal impact on the strength subtest, with values within the normal range by month 6, and a more distal impact on the running speed and agility subtest, with improvements evident at month 6 but not within the normal range until year 3.

**Handheld Dynamometry**

Handheld dynamometry (HHD) is a reliable and easy-to-use method to measure muscle strength [15]. In children and adolescents, torque can be calculated and compared with sex-specific norms [15,16]. In 5- to 12-year-old children with HPP, bilateral hip and knee extension and flexion, hip abduction, and grip strength were assessed by HHD [17]. Baseline strength ranged from median 32% (9.4, 52.7) predicted in the hip extensor to 60% (20.8, 149.2) predicted for grip (reported in torque for the right side as percent predicted for age- and weight-matched healthy peers). With asfotase alfa treatment, strength in all tested muscle groups, except grip, improved and continued to improve until the last assessment (P < 0.05); a median of 83% (45.7, 118.7) of predicted value was achieved for hip adductor at the last assessment (week 144 or 168) [17].

**6-Minute Walk Test**

The 6-minute walk test (6MWT) is a validated tool used to assess the distance a patient can walk on a level course in 6 minutes. Normative data are available for children and the distance walked can be compared as a percent of the predicted values by age, sex, and weight. The 6MWT reflects an integrated exercise response of multiple systems including the cardiorespiratory, neurological, and musculoskeletal systems and does not isolate the specific system of change. In 5- to 12-year-old children with HPP, a rapid improvement in 6MWT scores was demonstrated with asfotase alfa treatment: the median score increased from 61% predicted at baseline to within the normal range (80%–100% of predicted) after 3 months of treatment, and remained within the normal range through 5 years of treatment [18].

**Modified Performance-Oriented Mobility Assessment-Gait Characteristics**

The Performance-Oriented Mobility Assessment-Gait (MPOMA-G) to provide improved sensitivity for HPP-related impairments and to target gait as a disease-specific area of anticipated treatment benefit [19]. The modification process included validation with HPP key opinion leaders, evaluation of the relationship between the MPOMA-G scores and other clinical data points, and measurement of inter-rater reliability from archived gait videos in a natural history database. In a subset of a larger retrospective, noninterventional natural history study, all six children with HPP had clinically significant gait impairments that persisted, as assessed from gait videos using MPOMA-G [20]. In another study, all 14 children with HPP had gait impairments at baseline, and 8 patients treated with asfotase alfa showed significantly greater improvements compared with controls at the last assessment [21].
Conclusions

Pediatric rare diseases present unique challenges in clinical trial design and in selection of COAs that can be used to support claims in medical product labeling. Guidance is not available on best practices that deal with developing children, in whom cognition, motor skills, language, and level of independence in ADL vary greatly by age. In this example, multiple end points were required to capture multisystem impacts and to tell the complicated story from biochemical parameters to age-appropriate recreational and community participation. The measurement strategy that was used in HPP is optimal when the treatment-related improvements in symptoms and function are expected. When attenuation of disease progression is the treatment benefit or function is expected at a slower rate of skill acquisition than the typical population, AE scores or DQs should be considered.

Similar to HPP, many rare diseases present with multisystem impairments and a wide distribution of age and functional levels that are desirable to be included within labeling claims for medical product approval. It is imperative to consider multiple COAs early in the development process to design comprehensive prospective, observational, and natural history studies to gain insight into the multisystem impacts on age-appropriate markers. The HPP studies paired performance- and clinician-rated assessments to capture a quantitative measure of development and important caregiver perspective of the disease impact in the home community and school environment. The COAs that were highlighted for HPP are not necessarily specific to rare diseases. They are commonly administered in developmental evaluation and intervention clinics and reflect current standards for childhood assessment. COAs that provide normative data can help support market and payer approval and reimbursement for the approved drug intervention.

Source of financial support: There was no funding provided for this project or the associated article.

References