# CENTER FOR DRUG EVALUATION AND RESEARCH

***APPLICATION NUMBER:***

# 761052Orig1s000

# STATISTICAL REVIEW(S)

U.S. Department of Health and Human Services
Food and Drug Administration
Center for Drug Evaluation and Research
Office of Translational Sciences
Office of Biostatistics

# STATISTICAL REVIEW AND EVALUATION

BIOLOGICS LICENSE APPLICATION

| | |
|---|---|
| **NDA/BLA #:** | BLA 761052 |
| **Drug Name:** | Brineura (cerliponase alfa) Injection |
| **Indication(s):** | Late-Infantile Neuronal Ceroid Lipofuscinosis Type 2 (CLN2) |
| **Applicant:** | BioMarin Pharmaceuticals, Inc. |
| **Measure(s):** | Motor and Language Domains of the CLN2 Rating Scale |
| **Clinical Outcome Assessment (COA) Type:** | Clinician-reported Outcome (ClinRO) |
| **Date(s):** | Date Submitted: May 27, 2016<br>PDUFA Due Date: April 27, 2017<br>Review Completion Date: April 26, 2017 |
| **Review Priority:** | Priority |
| **Biometrics Division:** | DBIII |
| **Statistical Team:** | Min Min, PhD (Primary Statistical Reviewer)<br>Yeh-Fong Chen, PhD (Statistical Team Leader)<br>Lili Garrard, PhD (COA Statistical Reviewer)<br>Scott Komo, DrPH, Laura Lee Johnson, PhD (Concurrent Reviewers) |
| **Medical Division:** | DGIEP |
| **COA Staff Reviewer:** | Selena Daniels, PharmD |
| **Clinical Team:** | Elizabeth Hart, MD<br>Victor Baum, MD (Clinical Team Leader) |
| **Project Manager:** | Jenny Doan, MSN |

**Keywords:** BLA Review, Clinician-reported Outcome (ClinRO), scale comparability

## Table of Contents

# LIST OF TABLES



# LIST OF FIGURES

# 1 EXECUTIVE SUMMARY

The applicant has developed Brineura (cerliponase alfa) injection as an enzyme replacement therapy for the treatment of symptomatic pediatric patients three years of age and older with late-infantile neuronal ceroid lipofuscinosis (LINCL) type 2 (CLN2). Brineura (cerliponase alfa) is a drug-device combination product administered via a surgically implanted intraventricular (ICV) catheter. The applicant submitted data from a non-treatment natural history cohort (Study 190-901, hereafter referred to as Study 901) based on registry data; a phase 1/2, first-in-human, single-arm, open-label, dose-escalation Study 190-201 (hereafter referred to as Study 201); and the treatment extension Study 190-202 (hereafter referred to as Study 202). Study 901 included 42 evaluable patients (see Table 13). Twenty-four patients enrolled and 23 patients completed Study 201, with a dose escalation period of 30 mg, 100 mg, and 300 mg every other week, and a stable dose period of 300 mg every other week for a total of 48 weeks. Study 202 included 23 patients on a stable dose of 300 mg every other week for up to 160 weeks. A two-item version of the CLN2 rating scale (i.e., Motor and Language domains) was each used to collect the primary efficacy data of clinician-rated CLN2 scores in Study 901 and in Studies 201 and 202 (hereafter referred to as Study 201/202). The primary objective of Study 201/202 was to evaluate the safety, tolerability, pharmacokinetics, and efficacy of Brineura (cerliponase alfa) using the external natural history data as a comparator. The applicant's proposed primary efficacy endpoint was the proportion of patients with an absence of an unreversed (sustained) 2-point *rate* (slope) of decline or a score of 0 in the Motor-Language total score over 48 weeks. The Agency disagreed. Due to the issue of incomparability of measurements, the Agency focused on the Motor domain only. When the data were analyzed over 48 weeks, the efficacy findings were inconclusive. Since the study still was ongoing during the review, the Agency requested the applicant to conduct similar analyses for data over 72 weeks and also over 96 weeks, still focusing on the Motor domain only and for the matched population using the external natural history control.

Multiple issues and challenges were identified in comparing Study 201/202 data with Study 901 data. One major challenge was to determine whether the CLN2 rating scales (more discussion below) were comparable, since these two studies had different assessment times and were conducted by different methodologies (i.e., Study 201/202 had prospective assessments but Study 901 had both retrospective [parental recall interview and medical chart review] and prospective assessments). In addition, we noted that the assessment methods were different even within the same control subject over time. We found that the measurement properties of the CLN2 rating scale used in Study 901 could not be assessed; and the psychometric analyses conducted by the applicant using Study 201/202 data are limited in the evaluation of reliability (only inter-rater reliability could be assessed) and validity of the CLN2 rating scale (refer to Dr. Selena Daniels's review).

The Agency determined that the treatment and control groups used different CLN2 rating scales, specifically the use of different rater instructions for administration and training (using different anchor point definitions, i.e. descriptors for each response category) across studies (refer to Dr. Selena Daniels's review). However, because the case report forms (CRFs) from Study 201/202 included an identical form of the CLN2 rating scale (i.e., same anchor point definitions) to that

used in Study 901, an assessment of the comparability of the two CLN2 rating scales was performed. The Agency disagreed with the applicant's conclusion that adequate CLN2 rating scale similarity was demonstrated between the control and treatment studies. Based on the Agency's review of the CLN2 scale comparability video study, the scales used in Study 901 and Study 201/202 are not completely equivalent and have comparability issues, particularly with higher Language domain ratings by the Study 201/202 clinician (i.e., bias in favor of the treatment). The inconsistent Language domain ratings impede the interpretation and direct comparison of the applicant proposed Motor-Language total score within each study and across studies. As such, the Agency needs to have confidence that an observed change in CLN2 score is a real change and not due to measurement error by a 1-category decline. The majority of rating discrepancies observed in the video comparability study for the Motor domain were 1-category differences. Additionally, based on Dr. Selena Daniels's qualitative review, a score obtained from Study 901 may indicate a worse functional status than the same score obtained from Study 201/202 due to different anchor point definitions used to train raters, specifically for the score of 2. In summary, although the COA statistical reviewer replicated the applicant's CLN2 rating scale video comparability analyses, the COA statistical reviewer concludes that (1) the applicant submitted evidence is not sufficiently strong regarding the CLN2 rating scale comparability between Study 901 and Study 201/202; (2) the efficacy evaluation primarily should focus on the relatively more comparable Motor domain due to Language domain ratings being less comparable; and (3) a responder analysis using an absence of an unreversed (sustained) 2-*category* (raw) decline or an unreversed score of 0 in the Motor domain should be used as the primary analysis for evaluating the efficacy of Brineura (cerliponase alfa) in order to overcome the many measurement issues with the CLN2 rating identified in this BLA submission.

Before this BLA was submitted, the agency recommended the primary efficacy analysis be a responder analysis based on a matched population. The responder was defined as a patient who has an absence of an unreversed (sustained) 2-category (raw) decline or a score of 0 in the CLIN2 score. At that time, in the SAP no particular domain (either Motor or Language) was specified. Approximately four months after this BLA was submitted, the applicant informed the Agency that they identified some transcription errors in Study 901 data; they submitted the corrected Study 901 data one month later (11/2/2016). The applicant also submitted updated date of birth (DOB) for Study 201/202. The applicant's primary efficacy analysis was based on patients' rate of decline (slope) over 48 weeks, not the responder analysis the agency recommended. In addition, the agency's statistical reviewer noted that the applicant's matched population excluded one early terminated subject in Study 201. The primary statistical reviewer performed the McNemar's Exact test to analyze the matched 17 paired data for 48-weeks. To further confirm the efficacy findings by utilizing the entire patient population, after discussions with the clinical team 27 patients were excluded (see clinical review for details regarding the study inclusion criteria), the agency further performed a time to decline analysis, binary logistic regression analysis and ordinal analysis, in addition to the same matched analysis for 72-week data. The applicant also was asked to perform these analyses. In order to account for variable visit and data frequency in Study 901compared to Study 201/202's protocol of visits 8 weeks apart, each patient's last recorded response was imputed to the planned visits (i.e., 48, 72 and 96 weeks) for both Study 901 and Study 201/202 with the exception of the subject who terminated Study 201 early. The early terminated subject was marked as having a decline at the time of termination.

After numerous negotiations with the applicant and working together to establish a suitable final analysis plan and verified data for the evaluation of the Brineura's efficacy, the applicant submitted their results on March 24, 2017 and March 27, 2017. The primary statistical reviewer carefully reviewed their submission and was able to confirm their analysis results, including the matched analysis, ordinal analysis, time-to-decline analysis and binary logistic regression analysis based on 96-week data. The statistical review team concluded that the results based on 96-week data support the indication of Brineura (cerliponase alfa) to slow the loss of ambulation in symptomatic pediatric patients 3 years of age and older with late infantile neuronal ceroid lipofuscinosis type 2 (CLN2), also known as tripeptidyl peptidase 1 (TPP1) deficiency.

To further explore the extent of the study drug's efficacy, the primary statistical reviewer performed sensitivity analyses by imputing missing genotype information as different values. Those analysis results are supportive of the efficacy of Brineura (cerliponase alfa).

# 2 INTRODUCTION

## 2.1 Overview

The applicant, BioMarin Pharmaceuticals, is developing Brineura (cerliponase alfa) injection as an enzyme replacement therapy for the treatment of symptomatic pediatric patients three years of age and older with late-infantile neuronal ceroid lipofuscinosis (LINCL) type 2 (CLN2). CLN2 is a rare genetic disease and a form of Batten Disease characterized by the deficiency of tripeptidyl peptidase-1 (TPP1) enzyme, primarily affecting the central nervous system. In the US, the estimated incidence of CLN2 is approximately 0.5 per 100,000 births. Children typically are diagnosed around 2-4 years of age, and the disease results in seizures, ataxia, loss of motor skills, speech degeneration, blindness, and cognitive/developmental decline. CLN2 ultimately results in mid- to late childhood death. Currently there is no approved therapy for CLN2. Brineura (cerliponase alfa) is a drug-device combination product administered via a surgically implanted intraventricular catheter.

This licensing application includes data from a phase 1/2, first-in-human, single-arm, open-label, dose-escalation (Study 201) and the treatment extension (Study 202; see Table 1). To better understand the CLN2 disease patient population, a natural history cohort of 42 evaluable patients out of 69 patients (see Table 13) from a database put together by a clinical consortium (DEM-CHILD) was used. In addition, the applicant also proposed additional CLN2 disease natural history data from an ongoing prospective trial at Weill Cornell Medical College. However, at this time the review team has not fully assessed whether the Weill Cornell data can be considered as an appropriate external comparison to Study 201/202 (refer to Section 2.1.1).

This BLA submission serves two purposes: (1) to evaluate the safety, tolerability, pharmacokinetics, and efficacy of Brineura (cerliponase alfa), and (2) to assess the CLN2 rating scale adequacy and the CLN2 rating scale comparability between the natural history cohort study and the treatment study. This is a joint statistical review with analyses mainly conducted by both Dr. Min Min, the primary statistical reviewer and Dr. Lili Garrard, the Clinical Outcome

Assessment (COA) statistical reviewer for this BLA submission. Dr. Min Min's review focuses on the evaluation of efficacy for Brineura (cerliponase alfa). Dr. Lili Garrard's review focuses on the assessment of the CLN2 rating scale comparability between the control and treatment studies, which will precede the evaluation of efficacy for Brineura (cerliponase alfa) in this review.

For the evaluation of CLN2 rating scale adequacy, please refer to the review of Dr. Selena Daniels, the COA Staff reviewer for this BLA submission. Refer to the clinical review and division director summary for an evaluation of safety.

### 2.1.1 Clinical Studies Overview

The applicant submitted data from the non-treatment natural history cohort Study 901, the phase 1/2, first-in-human, single-arm, open-label, dose-escalation Study 201, and the treatment extension Study 202. Study 901 was based on the DEM-CHILD database, an independent consortium that collects and analyzes clinical, genetic, and biomarker data in patients with neuronal ceroid lipofuscinosis (NCL) diseases, including data on CLN2 patients since the 1960s. The purpose of Study 201 was to evaluate the safety, tolerability, pharmacokinetics, and efficacy of Brineura (cerliponase alfa) administered via a surgically implanted ICV catheter in patients with CLN2 disease. Study 202 was designed as a treatment extension study to evaluate the long-term safety and efficacy of Brineura (cerliponase alfa) in patients with CLN2 that had completed Study 201.

The applicant also proposed additional CLN2 disease natural history data from an ongoing prospective trial at Weill Cornell Medical College. Upon the Agency's request the applicant has submitted the Weill Cornell data, which consisted of 66 untreated subjects with a genotype diagnosis of CLN2 disease. However, few subjects had data for one year or longer. In addition, subjects from the Weill Cornell data were assessed with the Weill Cornell LINCL Scale (refer to Section 3.1.4), which is different from the CLN2 rating scale used for primary data collection in this BLA. Scale comparability between the CLN2 rating scale and the Weill Cornell LINCL Scale needs to be established should the applicant choose to use the Weill Cornell data as an external comparison to Study 201/202 (refer to Section 6.1). Additionally, at this time the review team has not thoroughly assessed whether the Weill Cornell data can be considered as an appropriate external comparison to Study 201/202 (e.g., comparable study populations).

An overview of the relevant clinical studies is presented in Table 1; and the Study 201 study design is depicted in Figure 1.

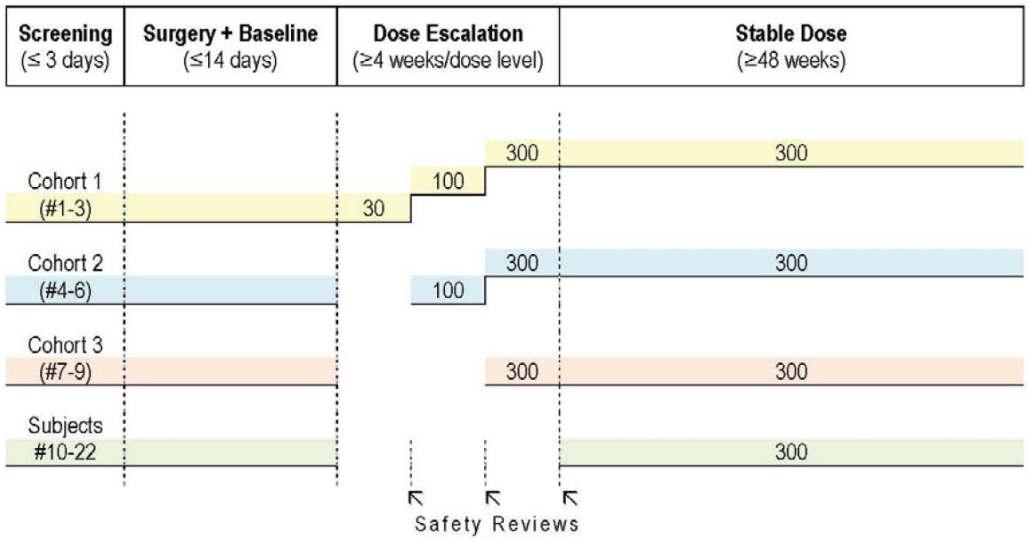
**Table 1: List of Relevant Clinical Studies**

| Study ID | Phase and Design | Study Population | Treatment Arm(s) | Number of Subjects | Duration |
|---|---|---|---|---|---|
| 190-901 | Non-treatment natural history control cohort based on registry data | Any child diagnosed with a type of neuronal ceroid lipofuscinosis (NCL; including CLN2) that has been confirmed through genetic testing | Do not apply | Overall: 69<br>Evaluable: 42 | Range: 2-61 months (based on data entered in DEM-CHILD database) |

| | | | | | |
|---|---|---|---|---|---|
| 190-201 | Phase 1/2, first-in-human, single-arm, open-label, dose-escalation | Children ≥3 years old with mild to moderate CLN2 disease, and a baseline Motor-Language summary score of ≥3 (with a score of at least 1 in each of the Motor and Language domains) | ICV infusion every 8 weeks:<br>• 30 mg<br>• 100 mg<br>• Stable dose 300 mg | Enrolled: 24<br>Completed: 23 | Stable dose treatment period: 48 weeks |
| 190-202 | Treatment extension study for subjects who completed 190-201 | | ICV infusion every 8 weeks:<br>• 300 mg | 23 | Stable dose treatment extension period: up to 240 weeks |

Source: COA Statistical Reviewer's table

**Figure 1: Study 201 Study Design**




Subjects 1-9 are assigned to three 3-subject cohorts to participate in the Dose Escalation Period. Once this period is completed, all subjects (including Subjects 10 through 22) are administered a stable dose of BMN 190 (300 mg or the highest dose tolerated) every two weeks for at least 48 weeks (Stable Dose Period). During the Dose Escalation period, cohorts are managed independently; enrollment in the next higher dose level follows a safety review by an independent Data Monitoring Committee. A dose level will not be recommended for further use if two or more subjects experience unacceptable toxicity at that dose level. Escalation of cohort starting dose will not be recommended if one or more subject experiences unacceptable toxicity.

Source: Applicant's Figure 9.1.1 of Protocol and Protocol Amendments 201.pdf of the BLA

## 2.2 Data Sources

The COA statistical reviewer evaluated the applicant's full evidence dossier and related datasets. The applicant's full evidence dossier is a psychometric report that includes the assessment of the CLN2 rating scale adequacy and the assessment of the CLN2 rating scale comparability between the natural history control study and the treatment study. This submission was submitted in eCTD format and was entirely electronic. The applicant's original BLA submission including the datasets is stored at the following location: \\CDSESUB1\evsprod\BLA761052\0001\m5. During the review, the Agency sent multiple information requests to the applicant. Please refer to Table 21 in Section 6.2 for a summary list of main COA related information requests, along with

locations of the applicant's response to these information requests. For efficacy datasets, please refer to Table 12 in Section 3.2.

# 3 STATISTICAL EVALUATION

## 3.1 Evaluation of CLN2 Rating Scale Comparability

This section of the review is conducted by the COA statistical reviewer and focuses on the assessment of the CLN2 rating scale comparability between the natural history cohort control Study 901 and the treatment Study 201/202.

### 3.1.1 COA Related Data and Analysis Quality

The COA statistical reviewer replicated the applicant's analyses to evaluate the measurement properties of the CLN2 rating scale and to assess the CLN2 rating scale comparability via a video study; however, the COA related data and analysis quality posed major challenges for the review team. The applicant's original BLA submission did not include sufficient information and data to provide supportive evidence for the CLN2 rating scale comparability between Study 901 and Study 201/202. The Agency had many iterations of communication with the applicant through information requests and additional teleconferences in order to gain a better understanding on many issues, e.g. detailed information regarding the DEM-CHILD registry data collection in which Study 901 data were based on, the data collection procedure and/or process in Study 201/202, the training and the method of rating within each study and across studies, the anchor point definitions used for the CLN2 rating scale administered in the control and treatment studies, the various methods used to collect the CLN2 rating scores, and the various versions of source worksheets and rating assessment guidelines (i.e. rater instructions for administration and training), etc. In addition, the Agency requested the applicant to submit corrected dataset for the scale comparability video study, all data tables, corrected analysis dataset and results for the CLN2 rating scale psychometric analyses, exact copies of all study instruments, additional scale comparability video analyses, corrected Weill Cornell data (not fully vetted at this time; refer to Section 2.1.1), etc. Please refer to Table 21 in Section 6.2 for a summary list of main COA related information requests.

The Agency acknowledged the applicant's efforts in responding to all the information requests, although the applicant's responses were not always adequate due to limitations and/or absence of evidence related to the external control data. Based on the totality of information, the overall data and the analysis quality of this BLA submission were determined to be acceptable (with documented limitations and/or absence of evidence) for assessing the CLN2 rating scale comparability between Study 901 and Study 201/202. For the analysis quality related to the CLN2 rating scale adequacy, please refer to the review of Dr. Selena Daniels, the COA Staff reviewer for this BLA submission.

### 3.1.2 CLN2 Rating Scale

The full-length version of the CLN2 rating scale is a clinician-reported outcome (ClinRO) measure that consists of four domains: Motor, Language, Visual, and Seizures. Table 2 below presents the anchor point definitions for each of the four domains. The CLN2 rating scale used in each of Study 901 and Study 201/202 is a two-item short-form version that aims to capture information on the Motor and Language age-equivalent functional domains of children with CLN2. Each domain can be scored on a 0 to 3 scale; and a Motor-Language total score (ranging from 0 to 6) is reported by summing the individual Motor domain score and Language domain score.

**Table 2: CLN2 Rating Scale—Full-Length Version**

| Hamburg Scale | | |
|---|---|---|
| Motor | 3 | Walks normally |
| | 2 | Frequent falls, obvious clumsiness |
| | 1 | No unaided walking or crawling only |
| | 0 | Immobile, mostly bedridden |
| Language | 3 | Normal |
| | 2 | Recognizably abnormal |
| | 1 | Hardly understandable |
| | 0 | Unintelligible or no language |
| Visual | 3 | Recognizes desirable object, grabs at it |
| | 2 | Grabbing for objects uncoordinated |
| | 1 | Reacts to light |
| | 0 | No reaction to visual stimuli |
| Seizures | 3 | No seizure in 3 months |
| | 2 | 1-2 seizures in 3 months |
| | 1 | 1 seizure per month |
| | 0 | >1 seizure per month |

From: Steinfeld, 2002, Am.J.Med.Genet.

Source: Applicant's Table 1.5.1.4.1 of Request of Breakthrough Therapy Designation.pdf (dated February 27, 2015)

It is important to note that the Agency determined that Study 901 and Study 201/202 used different CLN2 rating scales, specifically the use of different rating assessment guidelines across studies (refer to Dr. Selena Daniels's review). The rating assessment guidelines will be discussed below.

In Study 201/202, the CLN2 rating scale was administered at screening (to set entry criterion), baseline (prior to first infusion), on Day 4 after first infusion, every 4 weeks during the dose escalation period, at the start of the stable dose period, every 8 weeks thereafter, and at study completion/early termination visit. All in-clinic CLN2 rating scale assessments were videotaped.

It is important to note that each patient in Study 201/202 was rated by only one clinician (out of four; refer to Figure 2 in this review) at each assessment time point during the trial; and the clinician did not always rate the same patient throughout the trial. This approach to conduct the CLN2 rating assessments posed challenges for the evaluation of inter-rater reliability, which will be discussed in more details in Sections 3.1.5 and 3.1.6.

In the external control Study 901, CLN2 rating assessments were not conducted at regular intervals. Multiple methods (i.e., both prospective and retrospective [parental interviews and medical chart reviews]) were used to complete the external control CLN2 ratings with the majority of the CLN2 scores collected based on parental interviews with long recall periods and post-hoc review of medical records. The assessment method could differ even within the same control subject over time. On the other hand, Study 201/202 clinicians conducted prospective rating assessments for CLN2 patients based on direct observations of motor and language functions at the time of the assessment. Based on the applicant's response to the Agency's June 20, 2016 information request, "*testing [for language assessment] was conducted in the child's native language. If the evaluator was not fluent in the native language of the patient, a certified translator was utilized to discern if the vocalizations are interpretable language. A translator was necessary in 12 cases [out of 24 total patients].*"

Based on numerous iterations of communication with the applicant the Agency concluded that the clinical assessors from the control study and the treatment study were trained using different anchor point definitions. Table 3 shows the different versions of CLN2 rating assessment guidelines (RAG) used in Study 901 and Study 201/202. However, because Study 201/202 CRFs included an identical form of the CLN2 rating scale (i.e., same anchor point definitions) to that used in Study 901, an assessment of the comparability of the two CLN2 rating scales was performed (refer to Sections 3.1.5 and 3.1.6). For more details on the evaluation of the CLN2 rating scale and the RAG, please refer to the review of Dr. Selena Daniels.

**Table 3: CLN2 Rating Assessment Guidelines for Study 901 and Study 201/202**

| | Study 901 | | Study 201/202 | |
|---|---|---|---|---|
| *CLN2 Rating Assessment Guidelines (RAG)* | **Motor** | | | |
| | 3 | Walks normally | 3 | Grossly normal gait. No prominent ataxia, no pathologic falls. |
| | 2 | Frequent falls, clumsiness obvious | 2 | Independent gait, as defined by the ability to walk without support for 10 steps. Will have obvious instability, and may have intermittent falls. |
| | 1 | No unaided walking or crawling only | 1 | Requires external assistance to walk, or can crawl only. |
| | 0 | Immobile, mostly bedridden | 0 | Can no longer walk or crawl. |
| | **Language** | | | |
| | 3 | Normal | 3 | Apparently normal language. Intelligible and grossly age-appropriate. No decline noted yet. |
| | 2 | Has become recognizable abnormal | 2 | Language has become recognizably abnormal: some intelligible words, may form short sentences to convey concepts, requests, or needs. This score signifies a decline from a previous level of ability (from the individual maximum reached by the child). |
| | 1 | Hardly understandable | 1 | Hardly understandable. Few intelligible words. |
| | 0 | Unintelligible or no language | 0 | No intelligible words or vocalizations. |

Source: COA statistical reviewer's table (adapted from the COA Staff review)

### 3.1.3 Measurement Properties of CLN2 Rating Scale

The measurement properties of the Study 901 CLN2 rating scale could not be assessed due to the nature of the registry data and the lack of appropriate data elements. Documentation on the development of the CLN2 rating scale used in Study 901 is limited to Steinfeld et al. (2002).[1] Therefore, the applicant conducted psychometric evaluation for the CLN2 rating scale only using Study 201/202 data; and the results from the psychometric evaluation are included in the full evidence dossier submitted as part of this BLA. Although the applicant attempted to assess multiple measurement properties of the CLN2 rating scale, as pointed out by Dr. Selena Daniels's review, the psychometric analyses are limited in the evaluation of reliability (only inter-rater reliability could be assessed) and validity of the CLN2 rating scale. The construct validity of the CLN2 rating scale could not be fully assessed due to the limited selection of other appropriate Study 201/202 instruments and/or relevant domains within the other study instruments that measure similar concepts as the CLN2 rating scale. Furthermore, inter-rater reliability is most essential to the evaluation of the CLN2 rating scale comparability between Study 901 and Study 201/202. Therefore for this joint statistical review, the COA statistical reviewer focuses on detailed discussions on inter-rater reliability in Sections 3.1.5 and 3.1.6. For more discussion on the measurement properties of the CLN2 rating scale, please refer to the review of Dr. Selena Daniels.

[1] Steinfeld, R., Heim, P., von Gregory, H., Meyer, K., Ullrich, K., Goebel, H. H., & Kohlschütter, A. (2002). Late infantile neuronal ceroid lipofuscinosis: quantitative description of the clinical course in patients with CLN2 mutations. *American Journal of Medical Genetics Part A*, *112*(4), 347-354.

### 3.1.4 Other Study Instruments

In Study 201/202, five other instruments (three parent/caregiver-reported outcomes and two ClinROs) also were administered. The three parent/caregiver-reported instruments (i.e., Pediatric Quality of Life [PedsQL™] Generic Core Scales Version 4.0, PedsQL™ Family Impact Module Version 2.0, and CLN2 Quality of Life [CLN2QL] Questionnaire) were proposed by the applicant to help support the construct validity and responsiveness of the CLN2 rating scale. The Denver II Developmental Scale (a ClinRO) may be considered as an additional exploratory measure to help support the totality of evidence for assessing patients' motor function. The Denver scale has utility for detecting severe developmental problems, but it has limitations (i.e., unreliable) in predicting less severe or specific problems. In addition, the Denver scale is not a tool of final diagnosis, but can serve as a quick method to process large numbers of children in order to identify those that should be further evaluated. Due to concerns with these instruments' content, relevance, and limitation of interpretability, the COA statistical reviewer only briefly will present an overview of these instruments. Please refer to the review of Dr. Selena Daniels for a more detailed discussion on the three parent/caregiver-reported instruments.

As mentioned in Section 2.1.1, Weill Cornell LINCL Scale is a ClinRO measure used in the potential external Weill Cornell natural history data. The Weill Cornell LINCL Scale also was administered in Study 201/202. Since the review team has not determined the appropriateness of the Weill Cornell data as a potential external control, the Weill Cornell LINCL Scale will not be discussed in detail in this review.

*Parent/caregiver-reported outcome measures*
All three parent/caregiver-reported outcome measures have the same recall period, response options, scoring algorithm, assessment frequency, and similar requirement for translator assistance.

- Recall period: the past one month
- Response options: 0 "never," 1 "almost never," 2 "sometimes," 3 "often," and 4 "almost always"
- Scoring algorithm: each item is reverse-coded and linearly transformed to a 0-100 scale: 0=100, 1=75, 2=50, 3=25, 4=0. A mean score is calculated for each of the four domains only when at least 50% of the items in the domain are non-missing. If more than 50% of the items in a domain are missing, the mean domain score is not computed. A total mean score can be calculated as the sum of all items over the total number of items answered on all domains.
- Assessment frequency:
    - Study 201: baseline (prior to first infusion), start of the stable dose period, every 12 weeks thereafter, and study completion/early termination visit.
    - Study 202: every 24 weeks and study completion/early termination visit.
- Translator assistance: based on the applicant's response to the Agency's June 27, 2016 information request,
    - PedsQL™ Generic Core Scales Version 4.0 and PedsQL™ Family Impact Module Version 2.0: "*the UK English (at the UK, US and Italy sites) and German (at Hamburg site) language versions were used in Study 201. The parents/caregivers*

*fluent in English or German completed the appropriate self-report forms. For the parents/caregivers not fluent in one of these languages, the PedsQL™ Core and Family Impact Modules were administered by the site translator, who translated each item into the parent/caregiver's native language and recorded the responses on the appropriate language form (English or German) for the site's location.*"

- CLN2QL Questionnaire: "*the CLN2 QL was developed in English; the parents/caregivers fluent in English completed the English self-report form. For the parents/caregivers not fluent in English, the CLN2 QL was administered by the site translator, who translated each item into the parent/caregiver's native language and recorded the responses on the English language form.*"

PedsQL™ Generic Core Scales Version 4.0

The PedsQL™ Generic Core Scales Version 4.0 is a parent/caregiver-reported measure that contains 23 items that measure health-related quality of life in children and adolescents that are either healthy or with acute and chronic health conditions. Four domains are covered by the PedsQL™ Generic Core Scales instrument:

- Physical functioning (8 items)
- Emotional functioning (5 items)
- Social functioning (5 items)
- School functioning (5 items)

PedsQL™ Family Impact Module Version 2.0

The PedsQL™ Family Impact Module Version 2.0 is a parent/caregiver-reported measure that contains 36 items that measure the impact of pediatric chronic health conditions on parents and the family. Eight domains are covered by the PedsQL™ Family Impact Module instrument:

- Physical functioning (6 items)
- Emotional functioning (5 items)
- Social functioning (4 items)
- Cognitive functioning (5 items)
- Communication (3 items)
- Worry (5 items)
- Daily activities (3 items)
- Family relationships (5 items)

CLN2QL Questionnaire

Based on the applicant's response to the Agency's February 10, 2017 information request, "*the CLN2 disease-based quality of life (CLN2-QL) is an instrument developed by BioMarin. It was developed as a CLN2-specific health-related quality of life (HRQOL) instrument as there was no pre-existing tool available. As an exploratory endpoint, BioMarin recognizes that the instrument development process was not in full alignment with the 2009 PRO Guidance for Medical Product Development to Support Labeling Claims*." The CLN2QL Questionnaire contains 28 items that covers six disease-related domains:

- Seizures (6 items)
- Feeding (no G-tube) (4 items)
- Feeding (G-tube) (3 items)

- Sleep (5 items)
- Behavior (6 items)
- Daily activities (4 items)

*Clinician-reported outcome measures*

Denver II Developmental Scale

The Denver II Developmental Scale is a revision and update of the Denver Developmental Screening Test, designed to screen cognitive and behavioral problems and assess developmental milestones in infants and preschool children (from birth to six years). Proposed by the applicant as an exploratory study objective, the scale reflects what percentage of a certain age group is able to perform a certain task. The tasks are grouped into four domains:

- Personal social, e.g. smiling
- Fine motor adaptive, e.g. grasping and drawing
- Language, e.g. combining words
- Gross motor, e.g. walking

The Denver II Developmental Scale was administered at baseline (prior to first infusion), every 24 weeks thereafter, and study completion/early termination visit in Study 201; and every 24 weeks and study completion/early termination visit in Study 202. According to the applicant's response to the Agency's information request from the Late Cycle Meeting (dated February 21, 2017), only data from language and gross motor domains were collected in Study 201, as these two domains are the most relevant domains to the CLN2 Motor domain and Language domain scores. However, data from all four domains were collected in Study 202. As a result, data from personal social and fine motor domains are only available starting at Week 25 of Study 202.

Weill Cornell LINCL Scale

The Weill Cornell LINCL Scale has a similar design to the CLN2 rating scale, with the exception of different anchor point definitions. The Weill Cornell LINCL Scale also consists of two items that measure the Gait and the Language functional domains of children with CLN2. Similar to the CLN2 rating scale, each domain can be scored on a 0 to 3 scale. The Weill Cornell LINCL Scale anchor point definitions for both Gait domain and Language domain are presented in Table 4.

**Table 4: Weill Cornell LINCL Scale**

| Weill Cornell LINCL Scale (*as seen on source document and CRF*) | |
|---|---|
| **Gait** | |
| 3 | Normal |
| 2 | Abnormal but independent |
| 1 | Abnormal requiring assistance |
| 0 | Nonambulatory |
| **Language** | |
| 3 | Normal |
| 2 | Abnormal |
| 1 | Barely understandable |
| 0 | Unintelligible or no speech |

Source: COA statistical reviewer's table

The Weill Cornell LINCL Scale was administered together with the CLN2 rating scale at baseline (prior to first infusion), on Day 4 after first infusion, every 4 weeks during the dose escalation period, at the start of the stable dose period, every 8 weeks thereafter, and at study completion/early termination visit. All in-clinic CLN2 disease scale (both CLN2 rating scale and Weill Cornell LINCL Scale) evaluations were videotaped.

### 3.1.5 CLN2 Rating Scale Comparability Video Study

As mentioned in Section 3.1.2 (also refer to Section 6.1), the CLN2 rating scale comparability is needed to help establish evidence and confidence that the differences observed in the treatment and control studies are not due to differences with the CLN2 rating scale used in Study 901 and Study 201/202. As such, the Agency requested the applicant to submit a full evidence dossier including analyses to evaluate the CLN2 rating scale comparability between the natural history control and the treatment studies. The optimal and most robust approach recommended by the Agency was to rescore all the videotapes of Study 901 patients by the Study 201/202 clinicians, as this allows a more direct comparison of the findings from each study. However, the applicant indicated that this was not a feasible option based on previous communications with the Agency. In the submitted full evidence dossier, the applicant stated that "*due to the historical nature of this data, subsequent deaths of some of these children, and the fact that videotaping patients was not part of the clinical acquisition routine in Study 190-901, a sufficient supply of 190-901 patient videos was not available.*" Based on the applicant's response to the Agency's July 23, 2015 and August 7, 2015 information requests, a limited number of videos from Study 901 do exist, with approximately three to four videos (each from different disease stages) for two to four patients. However, the patient privacy protection laws in Germany are stringent and require re-consenting the patients for the provision and use of these videos. Moreover, the applicant was notified by the German site investigator that two patients were unwilling to consent to sharing their videos.

On the contrary, the treatment study protocols specified that all in-clinic CLN2 rating scale assessments were to be videotaped. Based on the applicant's response to the Agency's August 8, 2016 information request, 142 clinical assessment videotapes were produced across the multiple sites. According to the submitted full evidence dossier, each video of a clinical assessment was 30-90 minutes in length. The Agency previously recommended (in multiple correspondences with the applicant) that the scale comparability analyses should be conducted using all videotapes and multiple raters, as there might not be sufficient amounts of data needed to confirm comparability if utilizing only a single rater and only videos from a single site. However, due to language assessment issues and privacy laws, the applicant conducted analyses using only subsets of videos from the Hamburg, Germany clinical study site. Additional informed consent was obtained for this scale comparability video study.

Video Selection

As described in the submitted full evidence dossier, independent BioMarin staff with no direct involvement in Study 201/202 selected videos from the Hamburg site based on the following:

- Availability of Study 201/202 videos from baseline (prior to any dose) visit, Study 201 stable dose Week 25 visit, Study 201 completion visit, and Study 202 Week 25 visit (note that these assessment time points are study visits and not to be confused with nominal weeks. The corresponding nominal week for a particular study visit may differ for patients due to dose escalation);
- Confirmation that videos were taped on the day of Study 201/202 clinician rating and within the October 15, 2015 data cutoff date; and
- Review for quality and interpretability of language, as well as ensuring the redaction of any identifying or biasing information (e.g., audio of rater discussing scores). In the applicant's response to the Agency's July 20, 2016 information request, the applicant further clarified that review for quality and interpretability of language involved ensuring that the audio quality allowed for meaningful interpretation of patient language and that no technical problems with recordings would obstruct language assessments.

Rater Selection

Because each patient was rated by only a single clinician at each time point in Study 201/202, a direct assessment of the inter-rater reliability could not be performed. Given this limitation, Dr. (b) (6), was asked to rate selected videos from the Hamburg site in order to help provide an indirect assessment of inter-rater reliability of the CLN2 rating scale in Study 201/202.

In order to evaluate the CLN2 rating scale comparability between Study 901 and Study 201/202, the Agency requested the applicant to provide information on inter-rater reliability across both control and treatment studies as part of the full evidence dossier. The applicant selected (b) (6), who was not associated with the treatment Study 201/202, to rescore subsets of videos from the Hamburg site. According to the full evidence dossier, Dr. (b) (6) is a native German speaker and hence is "*capable of reviewing the language assessments on the videos that were conducted in or translated to the German language.*"

Video Study and Inter-Rater Reliability

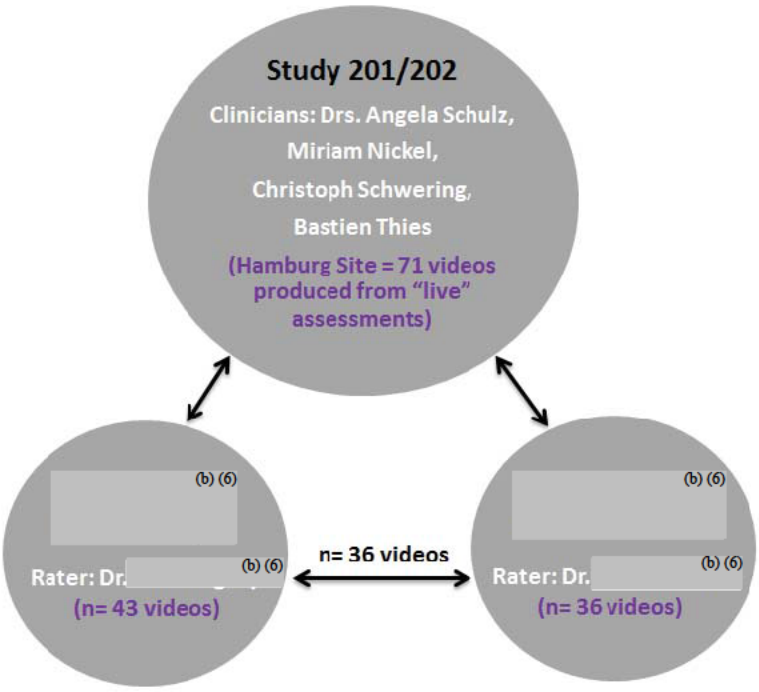
The video study was conducted based on eligible videos selected from a single site in Study 201/202, representing a total of 12 patients with a maximum of four assessment time points. The full evidence dossier stated that "*BioMarin provided the rater with videos without any information to identify the subject or the temporal sequence of the video, and the order of review was randomized to avoid bias.*" However, since there was no pre-specified sampling methodology, the selected convenience sample is potentially biased. The video study compared "live" clinician ratings with video assessment ratings by the two independent reviewers: the trainer of Study 201/202 clinician and the Study 901 CLN2 scale developer. It should be noted that in the submitted full evidence dossier, the applicant performed separate rater agreement assessments between the Study 201/202 clinician and the trainer of Study 201/202 clinician; and between the Study 201/202 clinician and the Study 901 scale developer. The Agency requested additional rater agreement analyses between the trainer of Study 201/202 clinician and the Study 901 scale developer as part of the July 20, 2016 information request. The additional analyses allow the Agency to have more evidentiary data to help further evaluate the scale comparability across studies, as both Study 201/202 trainer and Study 901 CLN2 developer conducted video assessment ratings. Any potential disagreements observed from this pair of raters would be attributed to other factors (e.g., differences in training), as opposed to other off-video information not captured on the videos.

Out of 142 total videos produced from in-clinic "live" assessments, 71 videos were from the Hamburg site. Among the 71 videos, the trainer of Study 201/202 clinician reviewed 36 videos; and the Study 901 CLN2 scale developer reviewed 45 videos. However, only 43 videos were included in the video study analysis. One video was excluded due to being taped several days after the patient's assessment and therefore did not include the clinician's assessment interview with the patient. Another video was excluded due to the assessment being performed after the initial October 15, 2015 data cutoff date of this BLA. The 36 videos reviewed by Dr. (b) (6), also were reviewed by Dr. (b) (6). Figure 2 provides a graphical presentation of the CLN2 rating scale comparability study.

The Agency acknowledges the challenge to evaluate inter-rater reliability and also the applicant's effort in producing contingency cross-classification tables and computing the weighted (i.e., quadratic weighting) Kappa statistic to demonstrate rater agreements. However, although helpful, the weighted Kappa statistic does not provide a full picture of rater agreement and discordance across all available time points for each patient included in the video study. Therefore, it is informative to conduct a graphical evaluation of rater agreement and discordance across all available time points for each patient, using a total of 36 videos reviewed by all three raters. In summary, the CLN2 rating scale comparability video study is conducted by examining the following:

- Graphical evaluation of rater agreement and discordance among all three raters
- Weighted Kappa and contingency tables
    - Study 201/202 clinician vs. Study 201/202 trainer
    - Study 201/202 clinician vs. Study 901 CLN2 developer
    - Study 901 CLN2 developer vs. Study 201/202 trainer

**Figure 2: CLN2 Rating Scale Comparability Study**




Source: COA statistical reviewer's figure

In order to help interpret the weighted Kappa statistic, the applicant specified in the full evidence dossier that the general guideline provided by Landis and Koch (1997) [2] is used to interpret the strength of agreement between different pairs of raters.

- < 0.00 = poor agreement
- 0.00 - 0.20 = slight agreement
- 0.21 - 0.40 = fair agreement
- 0.41 - 0.60 = moderate agreement
- 0.61 - 0.80 = substantial agreement
- 0.81 - 1.00 = almost perfect agreement

### 3.1.6 CLN2 Rating Scale Comparability Results and Conclusions

Weighted Kappa

Table 5 provides a summary of weighted Kappa statistic for the three comparison strata across all videos and by assessment time point, for the Motor domain, the Language domain, and the ML total score. The weighted Kappa results show that the Motor domain has a much higher rater agreement compared to the Language domain, as indicated by the substantial to almost perfect agreement across all videos, at each assessment time point, and across all pairs of raters. For the Language domain, although the Study 201/202 clinician had substantial to borderline almost perfect agreement with the trainer both at the overall level and at each time point, the clinician only had moderate agreement with the Study 901 CLN2 developer at the overall level. The

[2] Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 159-174.

clinician and Study 901 rater had fair agreement (weighted Kappa = 0.34) at Study 201 Completion and borderline substantial agreement (weighted Kappa = 0.62) at Study 202 Week 25. Similar observations can be seen with the Language domain ratings provided by the Study 201/202 trainer and the Study 901 rater.

The weighted Kappa results for the Motor-Language total score should be interpreted with caution as the higher weighted Kappa statistic is likely driven by the higher rater agreement observed in the Motor domain component of the total score.

**Table 5. Summary of Weighted Kappa Across All Videos and by Assessment Time Point**

| CLN2 Scale Comparability Study Comparison Strata | | | Motor Domain Weighted Kappa | Language Domain Weighted Kappa | ML Total Score Weighted Kappa |
|---|---|---|---|---|---|
| Study 201/202 clinician (via "live" assessment) | vs | Study 201/202 trainer (via ***36*** video assessments) | **Overall: 0.93**<br>Baseline: 0.76<br>201 Week 25: 1.00<br>201 Completion: 1.00<br>202 Week 25: 1.00 | **Overall: 0.82**<br>Baseline: 0.93<br>201 Week 25: 0.79<br>201 Completion: 0.67<br>202 Week 25: 0.80 | Overall: 0.92<br>Baseline: 0.92<br>201 Week 25: 0.93<br>201 Completion: 0.89<br>202 Week 25: 0.93 |
| Study 201/202 clinician (via "live" assessment) | vs | Study 901 CLN2 developer (via ***43*** video assessments) | **Overall: 0.88**<br>Baseline: 0.67<br>201 Week 25: 0.92<br>201 Completion: 1.00<br>202 Week 25: 0.90 | **Overall: 0.53**<br>Baseline: 0.57<br>201 Week 25: 0.55<br>201 Completion: 0.34<br>202 Week 25: 0.62 | Overall: 0.74<br>Baseline: 0.67<br>201 Week 25: 0.78<br>201 Completion: 0.69<br>202 Week 25: 0.79 |
| Study 901 CLN2 developer (via ***36*** video assessments) | vs | Study 201/202 trainer (via ***36*** video assessments) | **Overall: 0.94**<br>Baseline: 0.91<br>201 Week 25: 0.90<br>201 Completion: 1.00<br>202 Week 25: 1.00 | **Overall: 0.56**<br>Baseline: 0.59<br>201 Week 25: 0.50<br>201 Completion: 0.48<br>202 Week 25: 0.67 | Overall: 0.82<br>Baseline: 0.82<br>201 Week 25: 0.77<br>201 Completion: 0.80<br>202 Week 25: 0.88 |

Source: COA statistical reviewer's table

Graphical Evaluation of Rater Agreement and Discordance

Figure 3 shows 12 individual patient plots for the Language domain rating scores by the three raters, based on a total of 36 videos reviewed by all three raters. The x-axis represents the actual analysis day at each time point. The actual days are shown in red as the clinician's live assessment ratings are used for efficacy evaluation. The y-axis represents the Language domain ratings. Ideally, all three lines should be overlapping each other indicating perfect rater agreement. However, none of the patient plots has all three lines overlapping. Taking patient 1244-1004 as an example (located on the upper right corner of Figure 3), all three raters agreed on the first time point; on the second time point the Study 201/202 clinician agreed with the Study 201/202 trainer but disagreed with the Study 901 CLN2 developer; and all three raters

provided different scores on the last time point, resulting in a 2-category difference between the Study 201/202 clinician and the Study 901 CLN2 developer.

The submitted full evidence dossier stated that "*the [Study 201/202 "live" assessment] score was based on a comprehensive clinical assessment which may not be fully captured on the video*." In the applicant's response to the Agency's July 20, 2016 information request, the applicant provided clarification to this statement such that "*a rater-by-video would not have allowed evaluation of any possible interactions with the study clinician prior to the start of the videotaping (e.g., subject walks up to greet the study physician in the clinic hallway), any unobservable events (e.g., someone walking into the room), any pictures on opposite walls in the room or the view outside the window, etc., or any other unobservable details when viewing the video of an unknown stranger.*" The Agency acknowledges the applicant's rationale that the "live" assessments by the Study 201/202 clinician might have been influenced by other off-video interactions with the study subjects or their parents; however, this is unlikely the root of all the discrepancies and draws into question the comparability of the Language domain. To further support the Agency's concern regarding the comparability of Language domain, both the Study 201/202 trainer and Study 901 CLN2 developer provided ratings via video assessments; yet the two raters only had moderate rater agreement with a weighted Kappa statistic of 0.56 across all 36 videos (refer to Table 5 in this review). Therefore, the justification for "*a comprehensive clinical assessment which may not be fully captured on the video*" does not fully explain all the discrepancies observed with the video ratings.

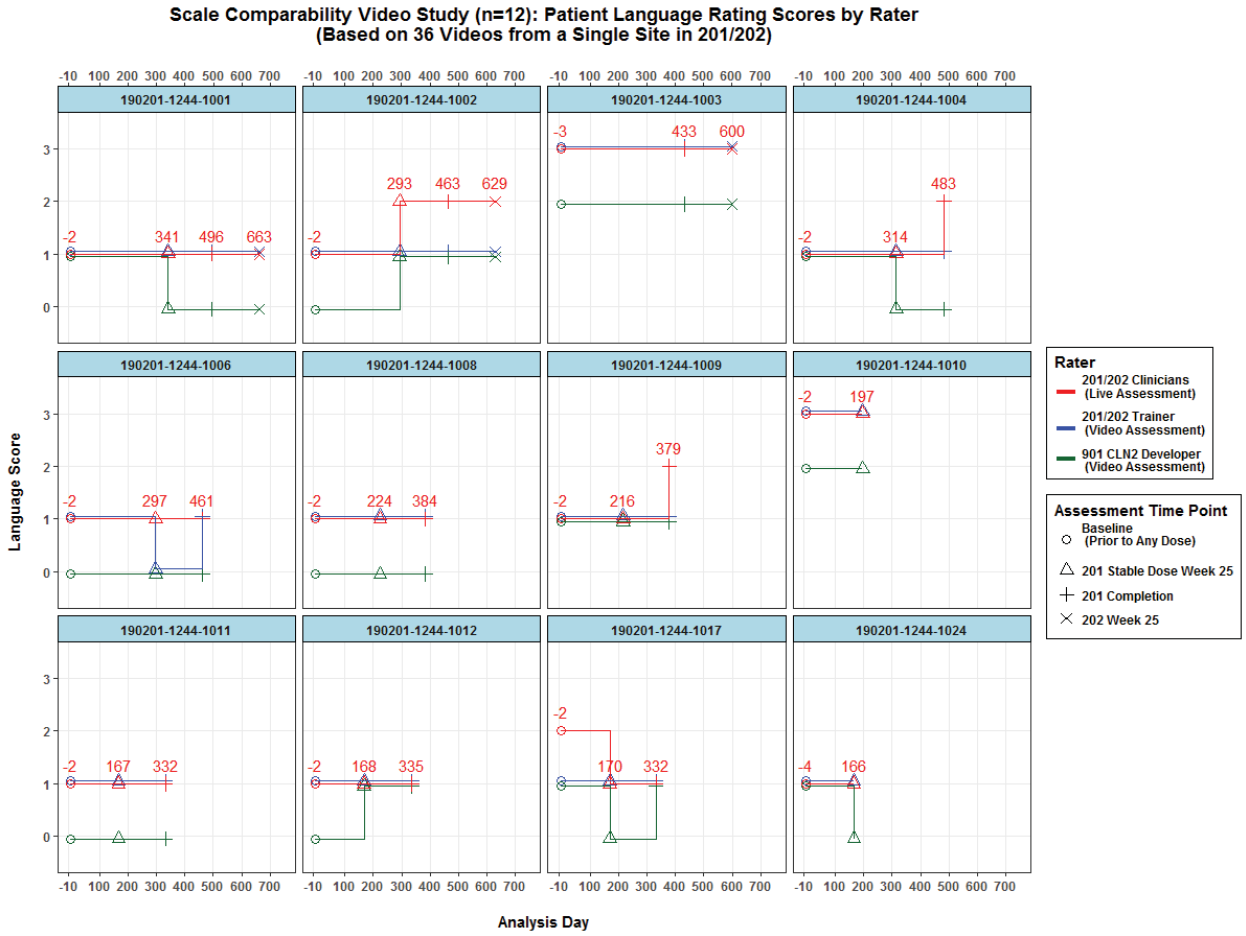
Based on all 12 patient graphs, it can be observed that although the Study 201/202 clinician did not always agree with the Study 201/202 trainer, there is a higher rater agreement between the clinician and the trainer (weighted Kappa = 0.82 across all 36 videos; refer to Table 5 in this review). On the other hand, the Study 201/202 clinician and the Study 901 rater had consistent rating discrepancies (weighted Kappa = 0.53 across all 43 videos; refer to Table 5 in this review) with higher Language ratings by the Study 201/202 clinician. Therefore, the COA statistical reviewer cannot state that the Language domain ratings are consistent between the external control Study 901 and the treatment Study 201/202.

Figure 4 shows the same 12 individual patient plots for the Motor domain ratings by the three raters. The Motor domain data show that the majority of patients had the same rating (with the exception of three patients) from all three raters across different time points. Results show that although the Motor domain scores are not perfectly replicated by all three raters, there is clearly a higher rater agreement compared to the Language domain scores, with an overall weighted Kappa of 0.88 between the Study 201/202 clinician and the Study 901 CLN2 developer.

Figure 5 shows the 12 individual patient plots for the Motor-Language total rating scores by the three raters. Similar to the weighted Kappa results discussed previously, the Motor-Language

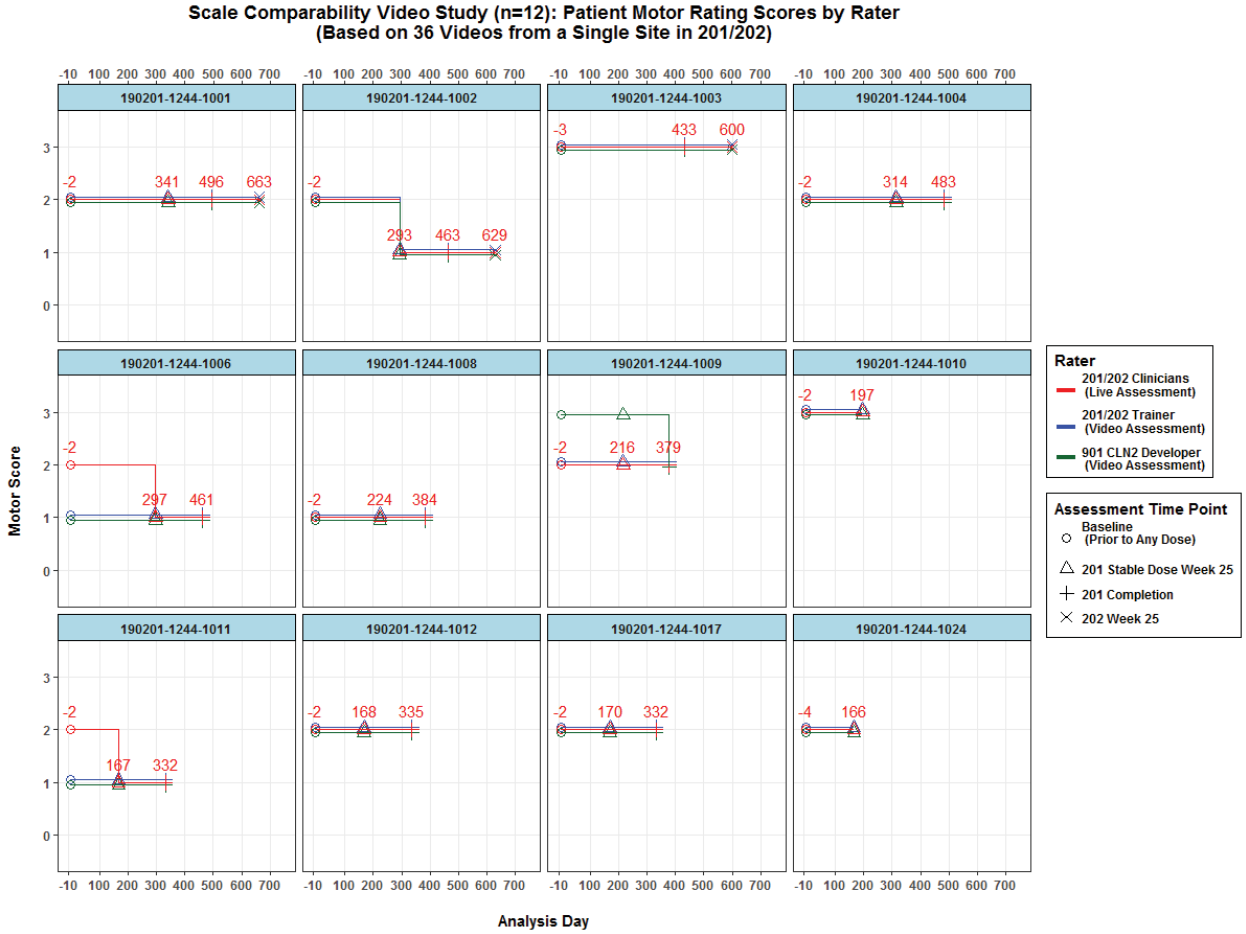
total score plots also should be interpreted with caution as the concerning comparability of the Language domain component impacts the ML total score.

**Figure 3: Patient Language Domain Rating Scores by Rater**



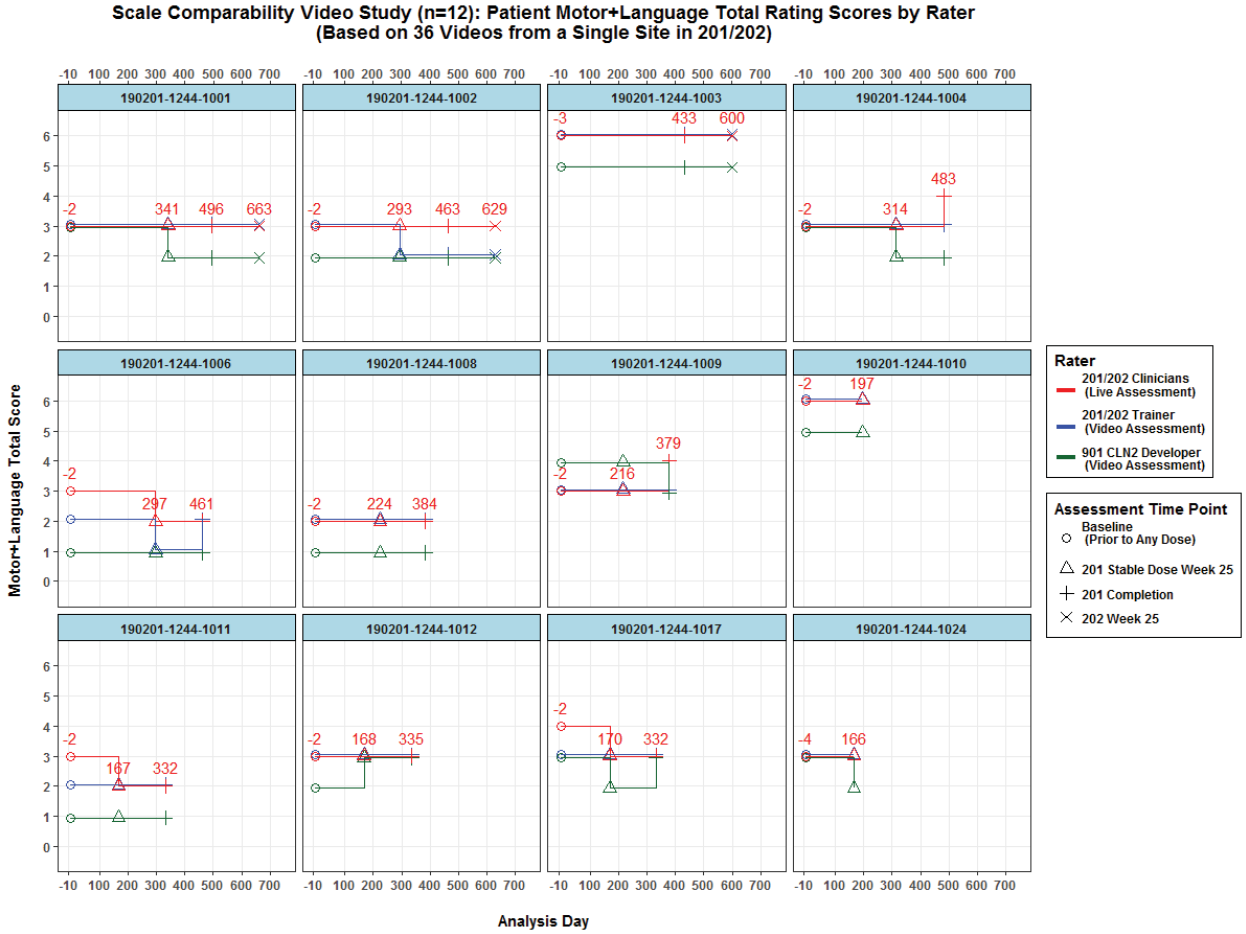

Source: COA statistical reviewer's figure

## Figure 4: Patient Motor Domain Rating Scores by Rater




Source: COA statistical reviewer's figure

**Figure 5: Patient Motor-Language Total Rating Scores by Rater**




Source: COA statistical reviewer's figure

Contingency Cross-Classification Tables

The contingency cross-classification tables are produced to further support the previously discussed results from the graphical evaluation of rater agreement and discordance, and the weighted Kappa statistic. Tables 6-11 show the overall contingency tables of the Motor domain ratings and the Language domain ratings across all videos within each comparison stratum. The contingency tables of the ML total score are not reported in this review since the observed discrepancies in the ML total score contingency tables are similar to the observed discrepancies from the graphical evaluation of the ML total score, which were contributed by the inconsistent Language domain ratings. The number of video assessments with perfect agreement between a particular pair of raters is located on the diagonal of each of the contingency tables. Any rating discrepancy between a pair of raters is indicated by the red text. Similarly, the results show less severe rating discrepancies for the Motor domain compared to the Language domain, across all pairs of raters. The contingency cross-classification table results further support the conclusion that there are inconsistent ratings on the Language domain between the external control Study 901 and the treatment Study 201/202.

**Table 6. All Motor Videos: Study 201/202 Clinician vs. Study 201/202 Trainer**

| Study 201/202 Clinician | Study 201/202 Trainer | | | | Total |
|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | |
| **0** | 0 | 0 | 0 | 0 | 0 |
| **1** | 0 | 10 | 0 | 0 | 10 |
| **2** | 0 | 2 | 19 | 0 | 21 |
| **3** | 0 | 0 | 0 | 5 | 5 |
| **Total** | 0 | 12 | 19 | 5 | 36 |

Source: COA statistical reviewer's table (adapted from applicant's Table VI.B.1.b.i of the full-evidence-dossier.pdf of this BLA)

**Table 7. All Language Videos: Study 201/202 Clinician vs. Study 201/202 Trainer**

| Study 201/202 Clinician | Study 201/202 Trainer | | | | Total |
|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | |
| **0** | 0 | 0 | 0 | 0 | 0 |
| **1** | 1 | 24 | 0 | 0 | 25 |
| **2** | 0 | 6 | 0 | 0 | 6 |
| **3** | 0 | 0 | 0 | 5 | 5 |
| **Total** | 1 | 30 | 0 | 5 | 36 |

Source: COA statistical reviewer's table (adapted from applicant's Table VI.B.1.b.vi of the full-evidence-dossier.pdf of this BLA)

**Table 8. All Motor Videos: Study 201/202 Clinician vs. Study 901 CLN2 Developer**

| Study 201/202 Clinician | Study 901 CLN2 Developer | | | | Total |
|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | |
| **0** | 0 | 0 | 0 | 0 | 0 |
| **1** | 0 | 12 | 0 | 0 | 12 |
| **2** | 0 | 3 | 19 | 2 | 24 |
| **3** | 0 | 0 | 0 | 7 | 7 |
| **Total** | 0 | 15 | 19 | 9 | 43 |

Source: COA statistical reviewer's table (adapted from applicant's Table 2 of the full-evidence-dossier.pdf of this BLA)

**Table 9. All Language Videos: Study 201/202 Clinician vs. Study 901 CLN2 Developer**

| Study 201/202 Clinician | Study 901 CLN2 Developer | | | | Total |
|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | |
| 0 | 1 | 0 | 0 | 0 | 1 |
| 1 | 18 | 9 | 0 | 0 | 27 |
| 2 | 1 | 7 | 0 | 0 | 8 |
| 3 | 0 | 0 | 7 | 0 | 7 |
| Total | 20 | 16 | 7 | 0 | 43 |

Source: COA statistical reviewer's table (adapted from applicant's Table 7 of the full-evidence-dossier.pdf of this BLA)

**Table 10. All Motor Videos: Study 901 CLN2 Developer vs. Study 201/202 Trainer**

| Study 901 CLN2 Developer | Study 201/202 Trainer | | | | Total |
|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | |
| **0** | 0 | 0 | 0 | 0 | 0 |
| **1** | 0 | 12 | 0 | 0 | 12 |
| **2** | 0 | 0 | 17 | 0 | 17 |
| **3** | 0 | 0 | 2 | 5 | 7 |
| **Total** | 0 | 12 | 19 | 5 | 36 |

Source: COA statistical reviewer's table (adapted from Table 7 of applicant's response to July 20, 2016 information request)

**Table 11. All Language Videos: Study 901 CLN2 Developer vs. Study 201/202 Trainer**

| Study 901 CLN2 Developer | Study 201/202 Trainer | | | | Total |
|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | |
| 0 | 1 | 17 | 0 | 0 | 18 |
| 1 | 0 | 13 | 0 | 0 | 13 |
| 2 | 0 | 0 | 0 | 5 | 5 |
| 3 | 0 | 0 | 0 | 0 | 0 |
| Total | 1 | 30 | 0 | 5 | 36 |

Source: COA statistical reviewer's table (adapted from Table 9 of applicant's response to July 20, 2016 information request)

CLN2 Scale Comparability Study Conclusion

Although the COA statistical reviewer replicated the applicant's CLN2 scale comparability video study analyses, the COA statistical reviewer disagreed with the applicant's conclusion that adequate similarity was demonstrated between the Study 901 CLN2 rating scale and the Study 201/202 CLN2 rating scale. The COA statistical reviewer concludes that the applicant submitted

supportive evidence is not sufficiently strong regarding the CLN2 rating scale comparability between the external control Study 901 and the treatment Study 201/202. Given the many concerning measurement issues with the CLN2 rating scale used in Study 901 and Study 201/202, the Agency disagreed with the applicant's proposed responder analysis using an absence of an unreversed (sustained) 2-point *rate* (slope) of decline or a score of 0 in the Motor-Language total score over 48 weeks. A longer duration of the efficacy data is needed to examine the drug's effect at later time points (i.e., 96 weeks). Furthermore, because of the inconsistent scale ratings in the Language domain (i.e., higher Language ratings by the Study 201/202 clinician that are biased in favor of the treatment), the COA statistical reviewer concludes that the efficacy evaluation primarily should focus on the Motor domain, which has been shown to be more comparable across studies.

Prior to the BLA submission, the Agency proposed a responder definition of an absence of an unreversed (sustained) 2-*category* (raw) decline or a score of 0 for Motor and Language scores separately, in order to overcome the measurement issues based on the Agency's qualitative assessment of the CLN2 rating scale. Several considerations went into the Agency's proposal. The review team was concerned that due to different anchor point definitions used in Study 201/202 and Study 901, a patient could be rated as e.g. a 3 in one study and a 2 in another study (also refer to Dr. Selena Daniels's review). Similar issues could be raised with how patients would be rated using any of the four response categories. The many different measurement methods presented in the Study 901 data further complicate matters, as there is no information supporting the consistency of ratings among parental recall interview, medical chart review, and prospective in-clinic assessment. The review team requested a responder definition of an absence of an unreversed (sustained) 2-*category* (raw) decline or a score of 0 even for the relatively more comparable Motor domain in order to ensure that an observed change was an actual change and not due to measurement error. As later demonstrated in the video comparability study, measurement error by a 1-category decline is present in the Motor domain ratings. The majority of rating discrepancies observed in the Motor domain were 1-category differences. Additionally, as pointed out in Dr. Selena Daniels's review, due to different anchor point definitions used to train raters across studies, a rating score from the Study 901 CLN2 scale may indicate a worse functional status than the same rating score from the Study 201/202 CLN2 scale, specifically for the score of 2. Although responder analysis may have some limitations (e.g., loss of power), to overcome the numerous major measurement issues identified in this BLA submission, the COA statistical reviewer recommends responder analysis, using the responder definition described above, as the primary analysis for evaluating the efficacy of Brineura (cerliponase alfa). Additional ordinal analyses can be informative by treating the CLN2 ratings as ordinal; however, ordinal analyses will not address the CLN2 rating scale comparability issues within Study 901 and between Study 901 and Study 201/202. Please refer to Section 3.2 for a more in-depth evaluation of the efficacy of Brineura (cerliponase alfa) in the following sections.

### 3.2 Evaluation of Efficacy for Motor Scores

Due to the complexity of this BLA, we have sent over 30 information requests (IR) to the applicant and had three center director (CD) briefings. The detailed information for all statistics related IRs and their responses are summarized in the Table 22 in the Appendix. Table 12 summarizes the important IRs or meetings.

**Table 12. History of FDA Information Requests and Meetings during BLA Review**

| Date | Activity |
|---|---|
| **6/30/2016** | IR: Perform the 1:1 matching analysis based on baseline CLN2 score and age±3 months as well as CLN2 score, age±3 months and common genotype. When more than one match occurred the selection was narrowed further by matching on additional variables in the order (1) detailed genome, (2) sex, (3) country. The submitted matching analyses were all based on rate of decline and the population which excluded one early terminated 201 subject. |
| **8/29/2016** | IR: Updated efficacy data including an additional nine (9) months (data-cut in June 2016) were submitted on 09/02/2016.<br>\\cdsesub1\evsprod\BLA761052\0023\m5\datasets<br>Additional responder analyses (defined as absence of an unreversed 2 point decline) were requested based on change in M-L score over 84 weeks (also for motor and language score separately (defined as an absence of an unreversed 1 point decline) over 84 weeks). Furthermore, sensitivity analyses based on rate of decline (slope) were requested. |
| **10/26/2016** | A SAS transportable dataset for the 201/202 German subjects who had an imputed DOB in the original submission of your application with the updated DOB was submitted. (\\cdsesub1\evsprod\BLA761052\0046\m5\datasets) |
| **11/02/2016** | IR: There were transcription errors for the 901 dataset originally submitted for BLA submission. **Corrected 901 datasets submitted (SDTM and ADAM); ISE datasets based on corrected 901 (N=42) and updated 202 data for DOB:**<br>\\cdsesub1\evsprod\BLA761052\0050\m5\datasets |
| **11/23/2016** | Datasets, SAS programs and matching analysis results for motor only submitted<br>\\cdsesub1\evsprod\BLA761052\0061\m5\datasets\ise\analysis\adam |
| **12/15/2016** | First center director (CD) briefing (48-week data): no substantial evidence of efficacy was found. |
| **01/13/2017** | IR: Matching analysis results and time to event analysis results for Motor-Language only submitted (48 and 72 weeks)<br>Matching datasets, define file, reviewer's guide and SAS programs submitted<br>\\CDSESUB1\evsprod\BLA761052\0080\m5\datasets\ise\analysis\adam |
| **01/27/2017** | Second center director (CD) briefing (72-week data): no substantial evidence of efficacy was found. |
| **01/31/2016** | IR: CORNELL data submitted (requested on 9/1/2016)<br>\\CDSESUB1\evsprod\BLA761052\0089\m5\datasets\190-901-supplement\tabulations\legacy\datasets |
| **02/16/2017** | IR: Efficacy analysis results using these 96 week data, the corresponding datasets (42/22), SAS programs, define file and reviewer's guide submitted<br>\\CDSESUB1\evsprod\BLA761052\0096\m5\datasets |
| **02/21/2017** | Teleconference with the applicant for analysis plan |

| | |
|---|---|
| **03/09/2016** | Supplemental SAP submitted \\CDSESUB1\evsprod\BLA761052\0107\m1\us\111-information-amendment\1113-efficacy-information-amendment |
| **03/15/2017** | IR: Response to FDA comments on supplemental SAP \\CDSESUB1\evsprod\BLA761052\0116\m1\us\111-information-amendment\1113-efficacy-information-amendment |
| **03/22/2017** | Teleconferencewith the applicant for supplemental SAP |
| **04/18/2017** | Third center director (CD) briefing (96-week data): full approval recommeded for the study drug |

Source: the primary statistical reviewer's table

### 3.2.1 Data and Analysis Quality

For the review team, there were many challenges in this BLA review. There were two external controls (Cornell data and Study 901) proposed by the applicant.

The major challenges are listed below for efficacy related data quality:

- Upon the Agency's request (9/1/2016) the applicant submitted the Weill Cornell data (1/31/2017). The review team has not thoroughly assessed whether the Weill Cornell data can be considered as an appropriate external comparison to Study 201/202 (e.g., comparable study populations).

- The score assessments were done every 8 weeks in Study 201/202. Study 901 is a natural history registry rather than a clinical trial; patients were not required to have clinic visits at specific intervals but as deemed necessary by the investigator for a specific patient. Therefore, intervals between clinic visits vary a lot in Study 901 (see Section 3.1.2).

- Corrected Study 901 data was submitted on 11/02/2016. Transcription errors for the natural history cohort Study 901 were found on 09/29/2016. The applicant claimed that the originally submitted Study 901 and the corrected Study 901 were similar because the population and disease progression as measured by CLN2 slope had not changed appreciably by comparing descriptive statistics between correct and incorrect Study 901 datasets. On 10/27/2016, the Agency requested the applicant to submit the corrected data and stated "We are still missing crucial data that will allow us to conduct a thorough review during this review cycle which should be submitted to the BLA by 11/01/2016."

- Updated Study 201/202 data with date of birth (DOB) was submitted on 11/02/2016: there were only birth years for twelve subjects from Study 201/202. The applicant imputed them as 6/30 for day and month and the Agency requested the applicant perform sensitivity analysis based on another imputed day and month as 12/1. The more precise revised date of birth was obtained by using the Denver II assessment where the CRF captures the age in years and months for the subject. An improved estimate of DOB was calculated by subtracting the age recorded at the Denver II assessment (in days) from the date of Denver II assessment. Age in days was calculated as the integer part of the quantity: (number of years x 365.25) + (number of months x 365.25/12). If there were

multiple assessments of the Denver II, then this calculation was produced for each of the assessments and the average computed as the estimate of DOB.

- Matching analyses results and data sets based on corrected Study 901 and updated Study 201/202 with DOB were submitted on 11/23/2016, 11/28/2016 and 12/2/2016 for Motor, Language, and Motor plus Language score, respectively. Furthermore, updated Study 201/202 raw data to 96 weeks (November 1, 2016 data cutoff), define file and reviewer's guide were submitted on 2/06/2017 and the analysis datasets were submitted on 2/16/2017.

- For all the recommended additional analyses or new data cut, no complete raw data, analysis data, define file and reviewer's guide were submitted. The agency had to request files repeatedly through IRs or teleconferences. Overall, it has been difficult and time consuming to locate all the datasets and the corresponding SAS programs needed to replicate the applicant's analysis results since they were scattered across many IR submissions after the agency emphasized the importance of submission of complete datasets at many communications.

For efficacy related analysis quality, throughout the BLA review, the agency had numerous iterations of communication with the applicant through IRs and teleconferences (see Table 13) in order to thoroughly evaluate the efficacy of Brineura (cerliponase alfa). In the original SAP (submitted on 9/21/2015), the applicant proposed to compare the mean rate of decline by using one sample t-test; we recommended to conduct responder analysis based on 2 scores decline for the matched population. The use of matched population was recommended due to the single arm, open-label trial design and an available natural history cohort. In the revised SAP, the primary endpoint was modified to identify responders based on rate of decline. At the t-cons (03/11/2016 and 5/19/2016), we clearly stated that the primary efficacy responder analysis needs to be based on a patient whose duration of any declining 2 scores is at least 12 months because only a few time points data would be available for each subject. Therefore, an individual subject's slope may not be meaningful. The applicant did not follow the Agency's recommendation; in the original BLA submission, the applicant's primary analyses were still the responder analyses considering individual patients' rate of decline (i.e., slope estimation).

In the ISE report from the original BLA submission, the applicant excluded one Study 201/202 subject due to early termination and only used Fisher's exact test, which did not incorporate the matched pair nature of the data. In addition, the matching window for baseline age was too wide (within 12 months) and the Agency recommended using a ±3-month window.

Per the COA statistical reviewer's conclusion regarding the incomparability of ratings for the Language domain between Study 201/202 and Study 901, the efficacy evaluation was performed primarily focused on the relatively more comparable Motor domain. The decision also was made to rely on the primary responder analysis using an absence of 0 or an unreversed 2-category (raw) decline in Motor score as the criterion for evaluating the efficacy of Brineura (cerliponase alfa) in order to overcome the many measurement issues with the CLN2 rating identified in this BLA submission. In particular, an unreversed 2-category decline means that any decline of 2-categories or more that had not reverted to a 1-category decline (or better) as of the last recorded

observation. An unreversed score of zero is a decline to 0 that had not reverted to >0 at the last recorded observation.

The major concerns and recommendations are briefly summarized in the following:

- According to the COA statistical reviewer's conclusion, the primary efficacy analysis is the responder analysis using an absence of 0 or an unreversed 2-category (raw) decline in Motor score.
- Analysis population needs to include one Study 201/202 subject due to early termination.
- Both Fisher's and McNemar's exact tests need to performed for all of the matched analyses.
- Because intervals between clinical visits vary a lot in Study 901, the agency recommended performing analyses using both the last available Motor score and next observation carried backward (NOCB) for the intermediate data points although the former one is determined as the primary.
- At the first CD briefing it was determined the matched analysis based on 48-week data did not provide substantial evidence to support the efficacy of Brineura (cerliponase alfa). An IR was sent on 12/23/2016 to request the Applicant conduct additional analyses: 1) additional matched analyses; 2) ordinal analysis (prepare an analysis plan and conduct analyses); 3) duration analysis (prepare an analysis plan and conduct analyses); and 4) develop a plan for Bayesian approach for Motor only based on 48-week and 72-week data. The applicant met with the Agency on 02/01/2017 to discuss the detailed plan for the re-analyses. Note that the Bayesian approach was not discussed nor planned.
- Matching analysis and time-to-event analysis results based on the 72-week data cut still did not provide substantial evidence (second CD briefing). During the teleconference (02/01/2017) with the applicant, the agency requested additional Study 201/202 data. The applicant offered a November 01, 2016 data cut (all subjects achieving 96 weeks of data) to support efficacy evaluation, and the agency agreed to review the additional efficacy data.
- On 03/09/2017, the applicant submitted a supplemental statistical analysis plan (SAP) based on the agency's recommendations. One important modification was that all of the following agreed re-analyses would be conducted by two types of patient populations (see Table 15). The agreed analyses are listed as follows:
    1. Matching analysis based on 96-week data;
    2. Binary logistic regression using all evaluable Study 201/202 and Study 901 subjects
    3. Ordinal analyses for 96-week data and in addition, including 48- and 72-week time points for repeated measurement analysis;
    4. Time to decline (defined as unreversed score of 0 or 2 category decline) analysis.

### 3.2.2 Efficacy Analysis Results

Table 13 shows that of the 69 subjects in the DEM-CHILD database, 42 were ultimately included in the evaluable population for Study 901.

**Table 13. Study 901: Patient Evaluability (DEM-CHILD Population)**

| | **Overall (N=69)** |
|---|---|
| Patients in the DEM-CHILD Population | 69 (100%) |
| Score assessment available | 60 ( 87%) |
| Patients who did not switch from 901 to 201/202 | 50 ( 72%) |
| One identical twin was excluded | 49 ( 71%) |
| With at least one score ≥6 months after first Motor-Language scale score | 42 ( 61%) |

Source: the primary statistical reviewer's table

Table 14 displays that in Study 201/202, in this review all the analysis results will be focused on N=22 population.

**Table 14. Study 201 Population**

| **Overall (N=24)** |
|---|
| Applicant originally excluded patient 1287-1007 from efficacy analyses due to early termination (N=23); the Agency included the patient in the efficacy analyses. The applicant later agreed to include the patient in analyses. |
| Exclude patients 1244-1010 and 1244-1003 who stay at the same ML score of 6 and thus were not considered to have motor or language symptoms at screening (N=22) (See Table 15, Population #1) |

Source: the primary statistical reviewer's table

Table 15 displays that two analysis populations were considered; in this review all the analysis results will be focused on Population #1.

**Table 15. Two Analysis Populations (Screening Baseline Used for Study 201/202)**

| | |
|---|---|
| **Population #1 (42/22)** | All subjects who entered the study with a baseline Motor/Language (ML) CLN2 scale score of 5 or less (N = 22) for Study 201/202; Study 901 baseline is defined as the time of the first CLN2 assessment at age ≥ 36 months and ML scale score < 6 |
| **Population #2 (42/24)** | All subjects who entered Study 201 (N = 24). Study 901 baseline is defined as the time of the first CLN2 assessment at age ≥ 36 months (regardless of ML scale score value). |

Source: the primary statistical reviewer's table

As mentioned in Section 3.1.2 (also refer to Section 6.1), the CLN2 rating scale comparability is needed to help establish evidence and confidence that the differences observed in the treatment and control studies are not due to differences with the CLN2 rating scale used in Study 901 and Study 201/202. As discussed in Sections 3.1.1-3.1.6, all the results in this section are based on Motor score only.

Since only responder analysis (responder defined as absence of an unreversed score of 0 or 2 category decline in CLN2 score over 48 weeks) was pre-specified, typically other analyses would only treated as post-hoc or sensitivity analyses. The agency recommended the primary efficacy endpoint be the proportion of patients with an absence of 0 or an unreversed 2-category (raw) decline in the Motor domain over 96 weeks, for the matched population using the external natural history cohort as a control and several other analyses to assess the totality of the evidence. Due to small sample size and the post-hoc nature of the analyses, no p-values are included in this section. Point estimates and their 95% confidence intervals (CI) are included.

Table 16 shows that there are two major differences in baseline characteristics between Study 901 and Study 201/202: 1) there are more males in Study 901; 2) all subjects in Study 201/202 were born after 2000 and 60% of Study 901 subjects were born before 2000. Genotype variations are discussed in Dr. Christine Hon's review.

**Table 16. Patient Disposition, Demographic and Baseline Characteristics**

| | Study 901 (n=42) | Study 201/202 (n=22) | Study 201/202 (n=24) |
|---|---|---|---|
| **Sex** | | | |
| Male | 25 (60%) | 7 (32%) | 9 (37.5%) |
| Female | 17 (40%) | 15 (68%) | 15 (62.5%) |
| **Genotype** | | | |
| 2 key mutations | 24 (57%) | 9 (41%) | 9 (38%) |
| 1 key mutation | 11 (26%) | 6 (27%) | 8 (33%) |
| No key Mutation | 7 (17%) | 7 (32%) | 7 (29%) |
| **Decade Born** | | | |
| Pre- 1980 | 4 (10%) | 0 | 0 |
| 1980s | 2 (5%) | 0 | 0 |
| 1990s | 19 (45%) | 0 | 0 |
| 2000s | 16 (38%) | 12 (55%) | 13 (54%) |
| ≥2010 | 1 (2%) | 10 (45%) | 11 (46%) |

Source: the primary statistical reviewer's results

Since the results based on Population #1 (42/22) and Population #2 (42/24) are very similar, in this section only the analysis results based on Population #1 were included.

Motor scores from the clinical study (i.e., Study 201/202) were compared to an independent natural history cohort (i.e., Study 901) that included 42 evaluable untreated patients. The natural history cohort follow up described below begins at 36 months of age or greater and at the first

time a Motor plus Language CLN2 score less than 6 was recorded. Based on this start time, 21 (50%) patients in the Study 901 experienced an unreversed (sustained) 2-category decline or unreversed score of 0 in the Motor domain of the CLN2 measure over a 96 week period and 5% (1 out of 22) subjects in Study 201/202 declined. In the following Tables 18-20 and Figure 6, "screening age" was defined in the Study 901 as the age at the first time a Motor plus Language CLN2 score less than 6 was recorded, and no earlier than 36 months of age. The "screening Motor score" of the natural history cohort was defined as the Motor score at the screening age.

Matched Analysis Results

Matching criteria (baseline Motor score, age±3 months, and genotype defined as 0, 1 or 2 key mutations) was used to match 22 Study 201/202 subjects with 42 Study 901 subjects. If there was a 1 to multiple or multiple to 1 match, further matching variables were considered in the following order: detailed genotype; gender; age of first symptom (looking at seizure first and if NA, then other symptoms). Table 17 shows the analysis results based 17 matched pairs at 48, 72 and 96 weeks.

**Table 17. Proportion of Patients**
**(Responder: Unreversed 2-category Decline or Score of Zero in Motor Domain)**

| | | 190-901 (Natural History) (n=17) | 190-201/202 (Brineura) (n=17) | Difference* | Odds Ratio** |
|---|---|---|---|---|---|
| | Time Point/Period | | | % (95% CI) | OR 95% CI |
| Response rate n (%) | Follow-up through Week 48 | 13 (76%) | 16 (94%) | 18% (-19, 51) | 0.25 (0.005, 2.53) |
| | Follow-up through Week 72 | 11 (65%) | 16 (94%) | 29% (-7, 61) | 0.17 (0.004, 1.37) |
| | Follow-up through Week 96 | 6 (35%) | 16 (94%) | 59% (24, 83) | 0.09 (0.002, 0.63) |

*confidence inernal for odds ratio based on binomial distribution
**confidence inernal for odds ratio based on McNemar's Exact test
Efficacy population based on full population minus two patients with baseline CLN2 score =6 (42/22)
Source: the primary statistical reviewer's table

Time to Decline Analysis Results

Given the non-randomized study design, a Cox Proportional Hazards Model adjusted for initial Motor score and genotype (0 key mutations (Y/N)) was used to evaluate time to unreversed 2-category decline or unreversed score of 0 in the Motor domain.

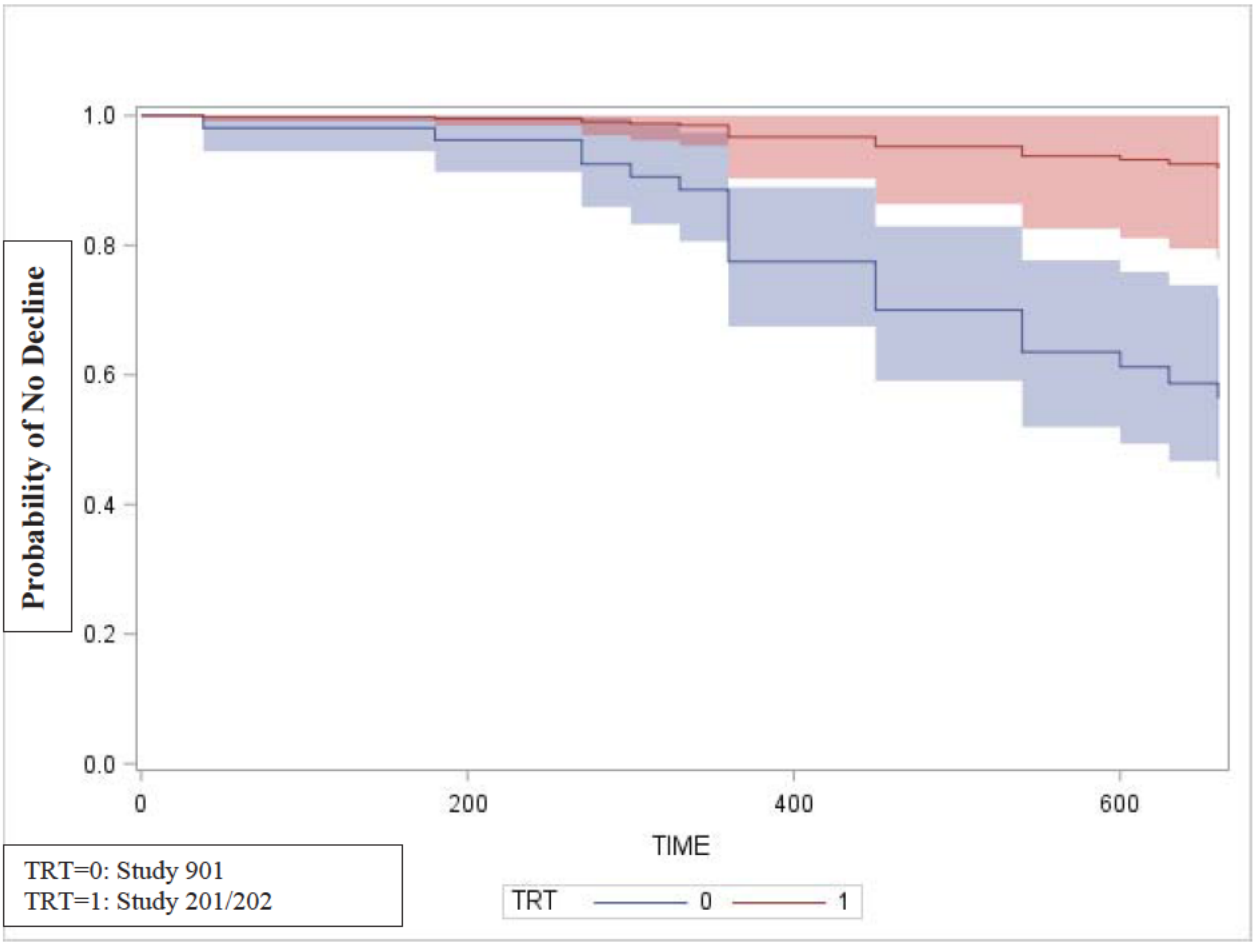
The applicant's result for time-to-decline analysis based on population #1 (42/22) using Cox-regression model is shown in Table 18.

**Table 18. Time-to-Unreversed 2-Category Decline or Unreversed Score of Zero in Motor Domain**
**(Follow up through 96 Weeks)**

| Model | Hazard Ratio | 95% CI |
|---|---|---|
| Covariates: screening baseline age, screening Motor score, and genotype (0 key mutations (Y/N)) | 0.141 | (0.02, 1.14) |

Source: the applicant's table 1.2 (March 27 2017 information request response)

**Figure 6. Estimated Time to Unreversed 2-category Decline or Score of Zero in Motor Domain for Symptomatic Pediatric Patients In the Brineura Single Arm Clinical Study 201 and its Extension Up to 96 Weeks Compared to Study 901**
**(Based on the Cox Proportional Hazards Model Adjusting for Covariates**




Hazard ratio: 0.12; 95% CI: (0.015, 0.92)
Study 901 follow up begins at 36 months of age or greater and at the first time a Motor plus Language CLN2 score less than 6 was recorded.
The Brineura treated population is the full population (N=24) minus the two patients with screening baseline Motor plus Language CLN2=6
Covariates: screening Motor score and genome: 0 key mutations.
Decline is defined as an unreversed (sustained) 2-point decline or unreversed score of 0 in the Motor domain of the CLN2 Clinical Rating Scale
Source: the primary statistical reviewer's figure

Ordinal Analysis Results (Week 96)
Ordinal nature of the Motor score defined as 0, 1, 2 and 3 needs to be considered. The applicant's result for ordinal analysis based on population #1 (42/22) assuming proportional odds is shown in Table 19.

**Table 19. Ordinal Analyses for Motor Score (Follow up through 96 Weeks)**

| Model | Odds Ratio | 95% CI |
|---|---|---|
| Covariates: screening baseline age and genotype (0 key mutations (Y/N)) | 0.170 | (0.05, 0.6) |

Source: the applicant's table 3.1 (March 27 2017 information request response)

Binary Logistic Regression Analysis Results (Week 96)
The applicant's result for binary regression analysis based on population #1 (42/22) at week 96 is shown in Table 20.

**Table 20. Binary Logistic Regression Analyses for Motor Score (Follow up through 96 Weeks)**

| Model | Odds Ratio | 95% CI |
|---|---|---|
| Covariates: screening baseline age, screening Motor score, and genotype (0 key mutations (Y/N)) | 0.08 | (0.007, 0.86) |

Source: adapted from the applicant's table 2.2 (March 27 2017 information request response)

### 3.2.3 Reviewer's Comments

1. **(Matching Analysis Results)** The statistical reviewer confirmed the applicant's matched analysis results. The matched population includes 17 pairs. Non-responder is defined as unreversed 2-point decline or unreversed score of zero on the Motor domain. For Population#1 (42/22), the response rate differences between treatment group (Study 201/202) and control group (Study 901) are 18%, 29% and 59% at Week 48, 72 and 96, respectively. Week-48 and -72 95% confidence intervals (CIs) for the odds ratio both includes 1; Week-96 95% CI excluded 1. Results for Population #2 were very similar and thus are not shown in this review.
2. **(Time to Decline Analysis Results for Follow-up Restricted ≤ 96 Weeks)** The statistical reviewer confirmed the applicant's time to decline analysis results based on 96-week data. The decline is defined as unreversed score of 0 or an unreversed 2 category decline in Motor score. Based on univariate analyses, genotype (0 key mutation (Y/N)) and screening Motor score are found to be significant covariates (see Table 23 in the Appendix). Since all the covariates did not violate proportional hazard assumptions, all the combinations of the significant covariates were explored using a Cox regression model. The AIC values ranged from 122.2 to 139 (see Table 24 in the Appendix) and

the smallest one was for the model including the covariates "*screening Motor score*" and "*genotype (0 key mutations (Y/N))*."

3. **(96-Week Ordinal Analysis Results)** The statistical reviewer confirmed the applicant's ordinal analysis results based on the 96-week data. All the combinations of the significant covariates were explored using logistic regression model. The AIC values ranged from 55.7 to 72 and the smallest one was found in the model including the covariates of "*screening Motor score*" and "*genotype (0 key mutations (Y/N))*". The upper bound of the 95% CI for the odds ratio excluded 1.
4. **(96-Week Binary Logistic Regression)** The statistical reviewer confirmed the applicant's binary logistic regression analysis results based on the 96-week data. According to the univariate analyses, the genotype (0 key mutations (Y/N)), birth year (<=2000 (Y/N)) and screening Motor score were found to be significant covariates. The final model included the covariates of "*genotype (0 key mutations (Y/N))*", "*screening age*" and "*screening Motor score*". The upper bound of 95% CI for the odds ratio excluded 1.
5. **(More Sensitivity Analysis Results for Genotype Covariate):** Two Study 901 subjects have missing genotype information. Two imputations were used (either 2 key mutations or less than 2 key mutations) and all the above analysis results are consistent.

### 3.3 Evaluation of Safety

The safety evaluation is not included in this review. Refer to the clinical review and division director summary for an evaluation of safety.

# 4 FINDINGS IN SPECIAL/SUBGROUP POPULATIONS

Due to small sample size, no subgroup analyses were performed.

# 5 SUMMARY AND CONCLUSIONS

### 5.1 Statistical Issues and Collective Evidence

The applicant submitted data from a non-treatment natural history cohort control Study 901 based on registry data; a phase 1/2, first-in-human, single-arm, open-label, dose-escalation Study 201; and the treatment extension Study 202. The primary objective of Study 201/202 was to evaluate the safety, tolerability, pharmacokinetics, and efficacy of Brineura (cerliponase alfa) using the external natural history data as a comparator. The applicant's proposed primary efficacy endpoint was the proportion of patients with an absence of an unreversed (sustained) 2-point *rate* (slope) of decline or a score of 0 in the Motor-Language total score over 48 weeks. The Agency disagreed with the applicant's proposal and instead recommended the primary efficacy endpoint to be the proportion of patients with an absence of an unreversed (sustained) 2-*category* (raw) decline or a score of 0 in the Motor domain over 96 weeks (to examine the drug's effect at later time points), for a matched population using the external natural history control. Due to the issue of incomparability of measurements, the Agency focused on Motor domain only. When the data were analyzed over 48 weeks, the efficacy findings were inconclusive.

Since the study still was ongoing during the review, the Agency requested the applicant conduct similar analyses for data over 72 weeks and also over 96 weeks, still focusing on Motor domain only and for the matched population using the external natural history control.

Multiple issues and challenges were identified in comparing Study 201/202 data with Study 901 data. One major challenge was to determine whether the CLN2 rating scales (more discussion below) were comparable, since these two studies had different assessment times and were conducted by different methodologies (i.e., Study 201/202 had prospective assessments but Study 901 had both retrospective [parental recall interview and medical chart review] and prospective assessments). In addition, we noted that the assessment methods were different even within the same control subject over time. We found that the measurement properties of the CLN2 rating scale used in Study 901 could not be assessed, and the psychometric analyses conducted by the applicant using Study 201/202 data are limited in the evaluation of reliability (only inter-rater reliability could be assessed) and validity of the CLN2 rating scale (refer to Dr. Selena Daniels's review).

The treatment and control groups used different CLN2 rating scales, specifically the use of different rater instructions for administration and training (using different anchor point definitions) across studies (refer to Dr. Selena Daniels's review). However, because Study 201/202 CRFs included an identical form of the CLN2 rating scale (i.e., same anchor point definitions) to that used in Study 901, an assessment of the comparability of the two CLN2 rating scales was performed. The CLN2 rating scale comparability video study conducted by the applicant used only a single rater and only subsets of videos from a single study site (due to language issues and privacy laws), whereas the Agency recommended using all videotapes from all sites and multiple raters. As demonstrated in the video comparability study, disagreements between ratings are noted on the scales used in Study 901 and Study 201/202 indicating that the scales are not completely equivalent due to comparability concerns with the Language domain. The higher Language ratings by the Study 201/202 clinician indicate bias in favor of the treatment. The inconsistent Language domain ratings impede the interpretation and direct comparison of the applicant proposed Motor-Language total score within each study and between studies. Due to the major measurement issues with the CLN2 rating scale, an absence of an unreversed (sustained) 2-*category* (raw) decline or an unreversed score of 0 was needed to ensure that an observed change was an actual change and not due to measurement error, where measurement error by a 1-category decline is present in the Motor domain ratings based on the video comparability study. The majority of rating discrepancies observed in the Motor domain were 1-category differences. Additionally, Dr. Selena Daniels's qualitative review points out that a score obtained from Study 901 may indicate a worse functional status than the same score obtained from Study 201/202 due to different anchor point definitions used to train raters, specifically for the score of 2.

In order to assess the efficacy of Brineura (cerliponase alfa), the responder analysis for 48-week data based on matched population was pre-specified before NDA submission. A responder was defined as a patient who had an absence of an unreversed (sustained) 2-category (raw) decline in CLN2 score over 48 weeks. In the SAP no particular domain was indicated. Since there was no substantial evidence of efficacy found in either the 48- or 72-week Motor data, the 96-week Motor data were mainly used to demonstrate the efficacy of Brineura (cerliponase alfa). The

responder definition was redefined as patients with an absence of an unreversed (sustained) 2-category (raw) decline or a score of 0 in the Motor domain over 96 weeks. Results of 96 week data demonstrated an upper bound of the 95% confidence interval (CI) for the odds ratio (based on 17 matched pairs) less than 1. The matching factors were screening Motor score, age (± 3 months) and genotype (0, 1 and 2 key mutations). In addition, since the overall population included 42 subjects from Study 901 and 22from Study 201/202, in order to maximize the use of available data, time to decline analysis, ordinal analysis (96 week) and binary logistic regression analyses were performed for this overall population. The upper bounds of the 95% CIs for the odds ratios are both lower than 1 for the binary logistic regression and ordinal analyses. The no decline rate difference is 59% between Study 901 and Study 201/202 data for Population #1 (42/22).

## 5.2 Conclusions and Recommendations

Although the COA statistical reviewer replicated the applicant's CLN2 rating scale video comparability analyses, the COA statistical reviewer disagreed with the applicant's conclusion that adequate CLN2 rating scale similarity was demonstrated between the control and treatment studies. The COA statistical reviewer concluded that (1) there is a lack of sufficiently strong evidence to support the CLN2 rating scale comparability between the external control group (Study 901) and the treatment group (Study 201/202); (2) due to higher Language ratings by the Study 201/202 clinician, the efficacy evaluation primarily should focus on the Motor domain, which has been shown to be more comparable across studies; and (3) to overcome the numerous major measurement issues with the CLN2 rating scale, a responder analysis using an absence of an unreversed (sustained) 2-*category* (raw) decline or an unreversed score of 0 in the Motor domain should be used as the primary analysis for evaluating the efficacy of Brineura (cerliponase alfa).

All the analysis results based on 96-week data support the indication of Brineura (cerliponase alfa) to slow the loss of ambulation in symptomatic pediatric patients 3 years of age and older with late infantile neuronal ceroid lipofuscinosis type 2 (CLN2), also known as tripeptidyl peptidase 1 (TPP1) deficiency. To further assess the study drug's efficacy by exploring the extent of the efficacy, the primary statistical reviewer performed sensitivity analyses by imputing missing genotype information as different values. Those analysis results are supportive of the efficacy of the study drug.

# 6 APPENDIX

## 6.1 Regulatory History

The applicant's development program for Brineura (cerliponase alfa) was designated as an orphan drug on April 1, 2013 and received Breakthrough Therapy Designation (BTD) on August 27, 2015 (based on a second BTD request from the applicant).

On February 27, 2015, the applicant submitted an initial request for BTD that subsequently was denied by the Agency after a careful review of the applicant's BTD request submission. In the BTD request submission, the applicant included the natural history cohort Study 901 and the phase 1/2, open-label, dose-escalation Study 201. The proposed primary endpoint used to quantify the CLN2 disease progression was the aggregate (sum) of the Motor and Language domains from the CLN2 rating scale.

On April 27, 2015, the Agency issued a BTD Request Denial letter and stated in the letter several limitations of the rating scale used to obtain the clinical data presented in the February 27, 2015 BTD request:

- *"The rating scales used to assess treatment effect in prospectively treated BMN 190 [cerliponase alfa] subjects vs. natural history controls are not the same, making cross study comparisons difficult to interpret. In the scale used for the ongoing study, the domain elements have been significantly modified (two domains were left out and the remaining 2 domains used slightly different definitions). Furthermore, you have not provided a rationale for the modifications made.*
- *Outcomes obtained from interviews of parents, relying on their recall, and the absence of structured assessments raise questions about the validity of the reported outcomes. In addition, in an open label treatment setting, lack of blinding could bias parental assessments, particularly when relying on recall. The open label setting, coupled with the nature of the assessment measures used, render comparisons to historical controls difficult to interpret.*
- *Given the international nature of the clinical development program, and in consideration of issues of cultural adaptation and translatability, it is unclear if either rating scale for 'language'/speech is accurate or reliable for parental report among subjects who speak a variety of primary languages.*
- *It is not clear if the rating scale used in the ongoing study was able to detect clinically meaningful changes or deterioration. The instrument used to measure clinical deterioration may not be reliable as content validity, inter-observer, and intra-observer reliability have not been established."*

An additional limitation of data was stated in the letter as following:

> *"The data for 9 subjects treated with BMN-190 reflect a short duration (9-12 months) of exposure. The analytical methods used to compare these preliminary data to historical controls do not provide preliminary clinical evidence that treatment with BMN-190 may demonstrate substantial improvement. Our review of the natural history data does not indicate a clear decline over the course of 9-12 months, leading us to conclude that a longer duration of follow-up may be needed to detect a substantial improvement between subjects treated with BMN-190 and natural history subjects."*

The Agency also made the following recommendations should the applicant submit a new BTD request:

- *"Provide clinical endpoint data obtained using a tool to measure change that yields reliable and interpretable data from both BMN-190- treated subjects and natural history controls. Additional prospective control data may be available from the ongoing Cornell study, which would make it possible to apply a rating scale that provided reliable and interpretable language and gait data in both a BMN-190-treated and an untreated population. In addition, given that you have videotaped all clinician rater assessments of gait in subjects on BMN-190, the interval gait assessments could be rescored in a blinded fashion using a scoring scale that is deemed to be sensitive to change and reflects attention to measurement principles.*
- *Because the natural history data you provided to us suggest that a longer duration study is needed to detect a substantial improvement between subjects treated with BMN-190 and natural history subjects, provide additional follow-up clinical data of longer duration to support that there is a substantial improvement. Furthermore, at this time, you have only submitted interim data from 9 subjects, but you have enrolled 24 subjects. Submission of data from a larger number of subjects using an agreed upon clinical assessment method would be expected to help address issues in heterogeneity within the disease that could impede detection of substantial improvement relative to a natural history control group.*
- *Submission of the natural history datasets from your proposed natural history control group and from the ongoing Cornell study would enable FDA to reproduce your analyses, which may help support your position that the open label clinical data from BMN-190-treated subjects demonstrate a substantive change from what would be predicted in an untreated control group. FDA would like to discuss with you the application of a more rigorous analysis methodology for making comparisons to the natural history data (i.e., accounting for age and not adjusting or "time-shifting" the control data in the manner performed)."*

On May 08, 2015, the applicant submitted a point-by-point response to the Agency's BTD Request Denial letter (dated April 27, 2015) as part of the meeting briefing package for an upcoming meeting scheduled with the Agency on May 20, 2015. In response to the Agency's scale comparability concern between the treatment and historical control studies, the applicant stated that the scale used for the primary endpoint of Study 201 was adapted from the scale used in the historical control Study 901 to "*provide objective anchors to allow standardization in a multi-site study setting. These refinements were not considered significant modifications and were not expected to impact the interpretation of a major treatment benefit*." The applicant also notified the Agency that they had plans to continue analyze results from Study 201 and its associated extension Study 202 to generate longer term data.

On May 20, 2015, the Agency and the applicant met to discuss the overall development plan of Brineura (cerliponase alfa) for the treatment of patients with CLN2. The Agency commented on several concerns regarding the use of the Motor and Language domains from the adapted CLN2 scale as a primary efficacy endpoint. The Agency had the following concerns:

- Issues related to the implementation of the CLN2 scale across studies—the Agency sought clarification regarding how clinicians were trained and what forms or instructions the clinicians used to rate the children in each study. The Agency also asked the applicant to provide information regarding inter-rater reliability within Study 901, within Study 201, and across both studies. The Agency stated that "*if after reviewing this information we do not have convincing evidence that the implementation of the original and adapted Hamburg Scales [Study 901 CLN2 scale and Study 201/202 CLN2 scale, respectively] was sufficiently similar across the studies, a potential path forward would be to have clinician raters rescore the videotapes of the children from either the treatment study (Study 190-201) or Hamburg cohort (Study 190-901) using a common scoring system across the two studies, and then comparing the findings across both studies. We suggest that the version of the Hamburg Scale (original or adapted) that you use as the common scoring system is the one for which you have the most supportive evidence (e.g., scoring manual, inter-rater reliability analyses, etc.);*"
- Combining the Motor and Language domains into a single summary score if one domain happens to drive the combined score, while the other contributes only partially, the different contributions may not be acknowledged (and ultimately labeled) correctly.
- Adequacy of efficacy assessments obtained with the adapted CLN2 scale.
- *"Although the preliminary evidence submitted to date in this subgroup of 9 patients appears to suggest stabilization of neurological symptoms over the period of evaluation, a subgroup of 9 patients is relatively small."*
- *"An imbalance in the percentage of patients with the 622C>T genotype between Studies 190-201 and 190-901."*

During the meeting, the applicant provided some clarification around the changes made to the instrument used in Study 901 and implemented in the efficacy Study 201. The applicant agreed to provide more detailed responses to address the Agency's comments in an upcoming submission along with a new request for BTD.

On July 01, 2015, the applicant submitted a new BTD request (second request) along with a request for a Type C meeting (scheduled for July 29, 2015) with the Agency to further discuss the acceptability and comparability of the CLN2 rating scales used in Study 901 and Study 201. However, the July 29, 2015 meeting was postponed until September 15, 2015 upon the applicant's request to allow additional time to prepare replies to information requests from the Agency.

On July 16, 2015, the Agency sent an IR through email to ask the applicant to perform Mixed Effect Model for Repeated Measures (e.g. MMRM) analyses for total Motor-Language, motor and language score for each of Study 901 and Study 201, respectively. The applicant submitted a complete response to the IR by the week of August 17, 2015.

On August 27, 2015, the Agency determined that Brineura (cerliponase alfa) for the treatment of patient with CLN2 disease met the criteria for Breakthrough Therapy designation and the BTD was granted by the Agency based on matched analyses comparing each of the 9 treated subjects to natural history subjects matched for baseline CLN2 score and age when available as well as

MMRM analyses. In the BTD granted letter, the Agency stated that "None the less, based on this conservative analysis of the data presented, treatment with BMN-190 appears to slow decline and stabilize progression of verbal and motor decline in CLN2."

The scheduled September 15, 2015 Type C meeting with the Agency was canceled based on the applicant's request. The applicant notified the Agency that no further discussion was required after receipt of the preliminary comments from the Agency. In the preliminary comments, the Agency stated that "*based on the data that you have submitted to date, we are unable to conclude at this time that the scales [Study 901 CLN2 scale and Study 201/202 CLN2 scale] are adequately similar. As previously stated, rescoring videotapes of Study 190-901 patients with the Adapted 0 – 6 Hamburg scale [Study 201/202 CLN2 scale] for use in the final analysis for Study 190-201 is the optimal approach to establish comparability. Given that this is not possible, an alternative would be to re-score the videos in the 190-201/190-202 studies using the original Hamburg scoring scale [Study 901 CLN2 scale] so that an adequate bridge to our natural history data can be established. We recommend that assessors rescoring the 190-201/190-202 studies be unacquainted with the Adapted 0 – 6 Hamburg scale to avoid any bias. Although not a regulatory requirement, ideally the assessors would be one of the original Hamburg raters not involved in Studies 190-201/190-202.*" The Agency also encouraged the applicant to obtain additional natural history data given the modeling limitations of Study 901 data due to the small sample size. The Agency stated "*for example, longitudinal natural history data may be available from the Weill Cornell study from which you previously submitted cross-sectional data. If you choose to pursue comparisons between your clinical trial data and the Weill Cornell study data, these comparisons should be made separately from comparisons to the Hamburg data [Study 901 data]. Finally, you would need to establish comparability between the Weill Cornell scale and either the adapted Hamburg scale or the original Hamburg scale using similar approaches as described [previously].*"

Regarding efficacy analysis, the Agency stated "*also, responder analyses applying a more conservative assumption for meaningful change may help overcome potential differences in the rating scale, as well as following the patients for an adequate duration to demonstrate that the lack of change could not be accounted for by differences in the rating scale. From an efficacy perspective, the submission should include data that demonstrate a clear treatment effect; such data should overcome the methodological limitations of your trial design (i.e. lack of a concurrent control group, rating scale observations not blinded, substantial natural history data obtained retrospectively, differences between the scales used in the treatment and natural history studies). Specifically, the data submitted should distinguish a meaningful and sustained difference in the rate of motor and language deterioration in the BMN 190 treated subjects compared to control subjects using conservative assumptions (e.g. the mean time for natural history subjects to decline from a score of 5 to 3 was 20.6 months in your analysis using conservative assumptions).*"

On November 06, 2015, the applicant submitted a request seeking comments and advice from the Agency on the applicant's proposal to score patient videos from the ongoing Study 190-201/202 to establish a bridge to the Study 901 natural history data. The applicant stated in this written response request that the applicant accepted the Agency's recommendation (based on the September 15, 2015 meeting preliminary comments ) to utilize an original Study 901 assessor to

score patient videos from the ongoing Study 201/202 using the Study 901 CLN2 scale. The applicant also proposed to have the original Study 901 assessor review videotapes of three to four clinical assessments from 8-10 patients at the Hamburg, Germany site. The applicant believed that "*evaluation of this sample size is large enough…to provide a bridge to the natural history data…In addition, this proposal limits any potential reader fatigue or definition creep that may occur as a result of this exercise and allows for the assessor to evaluate the language subscale assessments conducted in either his native language (German) or with a translation to German*."

On January 04, 2016, in the IR the agency had the following important comments for integrated summary of efficacy (ISE) SAP submitted on September 21, 2015:

- *"We recommend the primary efficacy analysis be based on the responder analysis and the proposed mean decline rate comparison as a sensitivity analysis. A responder can be defined as a patient whose duration of any declining 2 scores is longer than 9 months (see your IR response to Question 4 on 08/07/2015). Furthermore, all of the previously conducted sensitivity analyses based on MMRM in 07/16/2015 IR should still be performed by using 2-month time interval.*
- *The primary analysis should be based on study 190-201/190-202 ITT population with matched subjects from study 190-901. Sensitivity analyses can be conducted for other analysis populations using study 190-201/190-202 and matched subjects from study 190-901."*

On January 11, 2016, the Agency provided a written response to the applicant and acknowledged the applicant's efforts to establish evidence of scale comparability between the control and treatment studies. The Agency agreed in general with the applicant's proposal to implement the CLN2 rating scale used in the natural history control study to assess subjects across several time points in Study 201/202. The Agency reiterated the critical importance of scale comparability such that the Agency could be confident that the differences observed in the treatment and control studies were not due to differences between the CLN2 rating scales. The Agency provided the following comments to the applicant's video analysis proposal: (1) rescore the videos from Study 201/202 across specific selected time points (including baseline and subsequent visits) that should span across at least 36 weeks; and the rater should receive the videos in a randomized order; (2) given that the Agency had concerns with rescoring only a subset of the videos might not provide enough information to conclude that the scales were comparable;, if only videos from 10 subjects were used, the selection of the 10 subjects should be at random to avoid any selection bias; and the Agency continued to strongly recommend using a higher sample size or the complete sample "*in order to confirm that changes in score are occurring at the same point in deterioration on both scales;*" and (3) provide subject level scoring information for each instrument and each video, and provide individual patient plots of score ratings from both instruments for each domain separately and by the Motor-Language total score. The Agency also suggested using more than one rater (blinded to original ratings) to rescore the videos from Study 201/202 from different sites, as "*restricting the number of raters and sites diminishes the robustness of this [scale comparability] analysis.*" Regarding responder analysis, the Agency stated "*The responder analysis is, in general, acceptable. However, we recommend you also perform a secondary analysis using a Cox Model. The sample size is small for an analysis using Cox but it will allow for variables to be placed in the model. We also*

*strongly suggest performing a responder analysis for each domain separately (i.e., Motor domain and Language domain) and by total score (motor-language aggregate score).”*

On January 27, 2016, the applicant submitted a background document in response to the Agency’s January 11, 2016 written response and in preparation of the upcoming teleconference with the Agency on March 11, 2016. The applicant stated in the background document that the scope of the originally proposed video analysis had been expanded to include (1) all 12 patients enrolled at the Hamburg, Germany site, (2) videos ranging from baseline through 48 weeks for all 12 subjects and through 72 weeks for 9 subjects, and (3) three to four time points per patient (i.e., baseline, week 24, week 48 [end of Study 201], and week 72 [week 24 of Study 202]). In addition, videos submitted to the assessor were randomized by visit; and the video rescoring would be conducted by the only eligible, original Study 901 CLN2 rating scale developer, as requested by the Agency in the September 15, 2015 meeting preliminary comments.

On March 11, 2016, the Agency and the applicant held a teleconference to discuss the applicant’s expanded video analysis proposal for demonstrating the CLN2 rating scale comparability between the natural history control and treatment studies. The Agency had the following important comments (among others):

- “*In addition to your proposed analyses, we would also like to see patient-level data (including score-level differences) for the Motor function domain and Language domain, separately, between the scales [Study 901 CLN2 scale and Study 201/202 CLN2 scale] at each time point. Of primary importance are the contingency cross-classification tables detailing the agreement and discordance between the two scales?*
- *To facilitate FDA assessment of the comparability of these two scales, please submit a full evidence dossier in support of the two scales with your BLA submission.*
- *We are concerned that video rescoring utilizing only a single rater and only videos from a single site will not provide sufficient amounts of data in order to confirm comparability.*
- *We acknowledge your concern with including additional rater(s). However, including additional raters may minimize the potential bias of using only one rater and allow you to evaluate some measurement properties of the [Study 201/202 CLN2] scale (e.g., inter-rater reliability) to give us an indication of the reliability of the scale.*
- *We recommend a trained certified translator be used for translations of the video reviews that include subjects who are not native German speakers.*
- *Include documentation for each assessment when clinician ratings and parental reports differed.*
- *As previously requested, please submit the videos of the clinical assessments for Study 190-201 and 190-202 with your BLA application.”*

During the meeting, the applicant committed that they would provide patient-level data for each domain separately and as a total score at each time-point, including score-level differences. The applicant also agreed to provide a full evidence dossier and submit all videos of the clinical assessments for Study 201/202. The Agency recommended that the applicant “*provide all evidence to date on the rating scales (including any clinician notes) in the evidence dossier and provide context and rationale for limited and absent evidence.*”

For SAP of ISE, the Agency had the following important comment:

*"No, we don't agree. You did not address our following recommendations stated in the Correspondence dated 01/04/16: 1) you need to demonstrate the goodness-of-matching for your studies; 2) the primary efficacy analysis needs to be based on the responder analysis (e.g., a patient whose duration of any declining 2 scores is at least 12 months) (see Question 1 response for sensitivity analyses), and; 3) the primary analysis should be based on Study 201/202 ITT population with matched study 901 subjects. Sensitivity analyses can be conducted for other analysis populations using study 201/202 and matched study 901 subjects. In addition, your SAP should include specific sensitivity analyses for dealing with missing data."*

On March 29, 2016, the Agency and the applicant met for a pre-BLA meeting. The Agency provided general comments on the BLA package submission. In addition, the Agency reiterated that the videos of the clinical assessments for Study 201/202 needed to be submitted with the BLA application and referred the applicant to the March 11, 2016 meeting comments. No further discussions regarding the CLN2 rating scale comparability video analysis occurred at the March 29, 2016 meeting. Furthermore, the Agency referred to March 11, 2016 meeting minutes for specific guidance on all of the statistical analyses, including the requested MMRM analysis. In addition to analysis datasets in the CDISC SDTM and ADAM format, the Agency stated: "*you should provide all of your analysis programs, including those for the primary, the secondary and exploratory analyses. Please also include the study protocols, SAPs, any related amendments, and regulatory correspondence between you and the agency for all of the studies.*"

On August 09, 2016, the Agency completed the filing review and granted priority review status for the BLA submission (dated May 27, 2016). However, the Agency received a major amendment to the BLA submission (i.e., submission of corrected natural history control data) from the applicant on August 29, 2016. Consequently, the Agency issued a review extension—major amendment letter on September 02, 2016 which extended the user fee goal date by three months to provide time for a full review of the submission.

## 6.2 Summary of Information Requests (IR)

**Table 21. Summary of Main COA Related Information Requests (IR)**

| IR Sent Date | IR Response Date | Major Issue(s) and IR Response Location |
|---|---|---|
| July 23, 2015 | July 31, 2015 | The Agency requested the applicant to provide evidence regarding the CLN2 rating scale comparability across control and treatment studies by recommending the optimal and most robust approach, which is to rescore all the videotapes of the Study 901 patients with the Study 201/202 CLN2 scale and use this data in the final analysis for Study 201.<br><br>\\\cdsesub1\evsprod\ind122472\0017\m1\us\111-information-amendment\1113-efficacy-information-amendment\response-to-req-for-info-23july2015.pdf |
| August 7, 2015 | August 21, 2015 | • Provide the number of videos available from Study 901.<br>• Of these available videos, provide the number of videos for different patients. |

| | | |
|---|---|---|
| | | • If there are more than one video for a particular patient, provide the time course for which these videos span.<br>• Provide the range of scores for the patients in this study.<br><br>\\cdsesub1\evsprod\ind122472\0023\m1\us\111-information-amendment\1113-efficacy-information-amendment\response-to-req-for-info-20aug2015.pdf |
| June 20, 2016 | June 27, 2016 | • For Study 901 indicate whether the motor and language scores were derived from parental report, post-hoc scoring based on a review of medical records, or prospectively obtained in clinic.<br>• Provide the primary language for each subject, the language used to evaluate the subject, and the primary language of the evaluator for each subject in the 190-201 study and if available for 190-901 study.<br><br>\\cdsesub1\evsprod\bla761052\0003\m1\us\111-information-amendment\1113-efficacy-information-amendment\response-clin-rfi-1.pdf |
| June 27, 2016 | July 6, 2016 | • Provide all data tables for the psychometric analyses performed for the Study 201/202 CLN2 scale in the Full Evidence Dossier.<br>• Provide an exact copy of all assessments used to evaluate construct validity of the CLN2 scale noted in the Full Evidence Dossier.<br><br>\\cdsesub1\evsprod\bla761052\0006\m1\us\111-information-amendment\1113-efficacy-information-amendment\response-clin-rfi-2.pdf |
| July 20, 2016 | July 27, 2016 | • For Studies 901-201 and 901-202, provide clarification and justification for the weighting scheme used for computing weighted Kappa agreement.<br>• Clarify whether study clinician ratings were pooled or not and describe the methodology used to pool the data.<br>• For the videos rated by both Dr. (b) (6) and Dr. (b) (6), repeat all analyses performed for the video study to compare ratings provided by Dr. (b) (6) and Dr. (b) (6).<br>• For Studies 901-201 and 901-202, provide clarification on what kinds of information may not be captured by the videos.<br><br>\\cdsesub1\evsprod\bla761052\0009\m1\us\111-information-amendment\1113-efficacy-information-amendment\response-clin-rfi.pdf |
| July 20. 2016 | August 15, 2016 | • Provide details regarding the DEM-CHILD data collection, source documentation and data entry processes to be obtained from the site in Verona, Italy.<br>• Review the registry and/or original source data for the natural history subjects and provide the following data:<br>a. How each CLN2 assessment was made (e.g. whether the data was derived from parental report, post-hoc scoring based on a review of medical records, or prospectively evaluated in clinic by a trained rater) in as many subjects during as many time points as possible.<br>b. The absolute number and percentage of total assessments, motor assessments with a score of 1 or 2, and language |

| | | |
|---|---|---|
| | | assessments with a score of 1 or 2 that were performed prospectively (and clarify if prospective CLN2 motor and language scores were performed consistently across sites)<br><br>\\cdsesub1\evsprod\bla761052\0015\m1\us\111-information-amendment\1113-efficacy-information-amendment\response-clin-rfi.pdf |
| August 8, 2016 | August 12, 2016 | • The definitions of the Gait subscale anchor points in the final version of the Rating Assessment Guide (dated 28 April 2014) differ from the initial version of the Rating Assessment Guide (dated 24 February 2014). Confirm which gait anchor point definitions was used by your clinician assessors in your clinical studies (i.e., confirm that a gait score 2, independent gait, was defined as the "ability to ambulate **10 meters** without help" prior to April 2014, and afterwards was defined as "ability to walk without support for **10 autonomous steps**."). If the initial version definitions were in fact used, clarify how 10 meters was determined.<br>• For Study 201/202, provide a list of video file names for all the clinical assessments at all time points that were videotaped.<br><br>\\cdsesub1\evsprod\bla761052\0014\m1\us\111-information-amendment\1113-efficacy-information-amendment\response-clin-rfi.pdf |
| August 17, 2016 | August 25, 2016 | • The case report forms (CRFs) provided for Study 201/202 indicate that the descriptors from the Study 901 CLN2 scale were used. Confirm if these CRFs were used for all subjects at all assessments in Study 201/202. Also, provide a rationale for why the descriptors from the Study 201/202 scale were not included in the CRFs.<br>• Clarify what processes were in place to ensure that the investigators used the Study 201/202 scale anchor definitions when responding to the CRFs.<br><br>\\cdsesub1\evsprod\bla761052\0019\m1\us\111-information-amendment\1113-efficacy-information-amendment\response-clin-rfi.pdf |
| August 24, 2016 (for the August 25, 2016 teleconference) | September 16, 2016 | • The Agency seeks to better understand how the Hamburg and Cornell CLN2 scores were obtained during studies 201 and 202. Describe the clinic study visit(s) with regards to:<br>a. History and examination procedures (e.g., specific instructions given to patient/caregiver to obtain CLN2 rating scores, specific instructions given to raters with regards to how to complete CLN2 scoring including document(s) to be reviewed prior to completing CLN2 scoring).<br>b. Procedure for completing the source worksheet.<br>c. Clarify which procedures were standardized in the protocol and/or training materials (and where this information can be found).<br>• Briefly describe version control and naming conventions that were used regarding versions and dates used for the Rating Assessment Guide and the worksheets. Update the sample table provided by the Agency that summarizes the documents (with the actual documents embedded) used by each site during the course |

| | | |
|---|---|---|
| | | of study 201 and study 202. Summarize the differences between documents and indicate which changes were required and what was left to the sites' discretion."<br><br>\\cdsesub1\evsprod\bla761052\0029\m1\us\111-information-amendment\1113-efficacy-information-amendment\response-clin-rfi.pdf |
| August 29, 2016 | September 8, 2016 | • Explain the scoring instructions, including definitions for motor scores 0-3 and language scores 0-3, used by Drs. Schulz and Nickel to score children who were prospectively evaluated in clinic. Indicate how frequently assessments at a single time-point were made by both doctors. Clarify if the doctors ever disagreed on the language or motor scores for children evaluated prospectively in the DEM-CHILDs registry.<br>• Clarify the instructions provided to Dr. Simonati and the Hamburg physicians so that language and motor CLN2 scores could be retrospectively assigned based on medical records. Clarify the scoring definitions for motor scores 0-3 and language scores 0-3, used by Dr. Simonati.<br><br>\\cdsesub1\evsprod\bla761052\0024\m1\us\111-information-amendment\1113-efficacy-information-amendment\response-clin-rfi.pdf |
| August 31, 2016 | September 8, 2016 | Clarify why a subject would have a different score for Hamburg motor and Cornell gait during a single clinic visit if raters are scoring subject using the Rating Assessment Guide in which the anchor point definitions are the same for Hamburg motor scores and Cornell gait scores.<br><br>\\cdsesub1\evsprod\bla761052\0025\m1\us\111-information-amendment\1113-efficacy-information-amendment\response-clin-rfi.pdf |
| September 1, 2016 | January 31, 2017 | The Agency requested Weill Cornell's study related documentation (e.g., Protocols, SOPs, patient data, publications, etc.), and available longitudinal, CLN2 motor-language data provided in a SAS dataset.<br><br>\\CDSESUB1\evsprod\BLA761052\0089 |
| September 14, 2016 (for teleconference) | September 22, 2016 | • Confirm that the table provided by the Agency accurately describes the methodology used for CLN2 scores for each subject in the 190-901 supplemental analysis. If you note any inaccuracies please correct and clarify.<br>• Has any standardization been performed for Study 901 data to determine reliability and comparability of non-0 CLN2 motor and language scores obtained from parental interviews compared to medical records compared to observations in clinic?<br>• The methodology section (page 348) of the 2002 Steinfeld et al. paper states, "loss of motor performance, seizure activity, loss of vision, loss of language was rated in such a way that the normal condition was given a score of 3, a slight or just noticeable abnormality a score of 2, a severe abnormality a score of 1, and a complete loss of function a score of 0...the scoring system was explained to the families, and scores were recorded during interviews of about 3 to 4 hr at their homes." Do you have additional information about the scoring performed for these |

| | | |
|---|---|---|
| | | patients and whether the patients were re-scored based on the scoring definitions that appear in the table in this publication?<br><br>\\cdsesub1\evsprod\bla761052\0032\m1\us\111-information-amendment\1113-efficacy-information-amendment\response-clin-rfi.pdf |
| November 15, 2016 | November 17, 2016 | Several data discrepancies were found in the analysis data of the video study report. Clarify all discrepancies discovered by the Agency (including any additional discrepancy not listed here but discovered during data checking), and submit a corrected video study SAS analysis data file.<br>• The full evidence dossier states that Dr. (b) (6) rated a total of 45 videos. However, the SAS data only contains 44 records for Dr. (b) (6).<br>• The full evidence dossier states that Dr (b) (6) rated a total of 36 videos for the inter-rater reliability analyses. However, the SAS data file does not match the line listing in Attachment 6 of the applicant's response to the Agency's July 20, 2016 information request.<br><br>\\cdsesub1\evsprod\bla761052\0058\m1\us\111-information-amendment\1113-efficacy-information-amendment\response-clin-rfi.pdf |
| November 18, 2016 | November 25, 2016 | The Agency was unable to duplicate some of the applicant's analyses results from the full evidence dossier for Construct Validity and Responsiveness. If any discrepancy is discovered during data checking, please identify the discrepancy and submit corrected analyses result(s).<br><br>\\CDSESUB1\evsprod\BLA761052\0062\m5 |
| December 1, 2016 | Part 1: December 2, 2016<br><br>Part 2: December 23, 2016 | The Agency would like to know when the applicant expects to be able to provide the assessment methods (by subjects and by time point) for the Corrected 901 DEM-CHILD data. The applicant has previously provided the Agency with a table and during the t-con on 10/27 and agreed to update this for the corrected DEM-CHILD data.<br><br>Part 1: \\cdsesub1\evsprod\bla761052\0065\m1\us\111-information-amendment\1113-efficacy-information-amendment\response-clin-rfi-01dec2016.pdf<br>Part 2: \\cdsesub1\evsprod\bla761052\0072\m1\us\111-information-amendment\1113-efficacy-information-amendment\response-clin-rfi-01dec2016.pdf |
| February 10, 2017 | February 16, 2017 | • Submit evidence of the development history of the CLN2 QOL, including its measurement properties (reliability, validity, ability to detect change), scoring information, and handling of missing data.<br>• Submit details on rater instructions and administration of assessments, including training materials.<br>• Provide information on what constitutes a meaningful change in the CLN2 QOL by domain and total score, if applicable.<br>• Provide information on whether parents/caregivers required the use of a translator or not when completing the CLN2 QOL, for each subject in Study 201/202. |

| | | |
|---|---|---|
| | | \\cdsesub1\evsprod\bla761052\0098\m1\us\111-information-amendment\1113-efficacy-information-amendment\response-clin-rfi.pdf |
| February 14, 2017 | March 3, 2017 | One subject is missing from the Cornell.xpt SAS data file. Please submit a corrected Cornell data file and clarify all discrepancies discovered during data checking.<br><br>\\CDSESUB1\evsprod\BLA761052\0106 |

Source: COA statistical reviewer's table

**Table 22. Summary of Main Statistic Related Information Requests (IR)**

| IR Sent date | IR Response date | Major Issue(s) and IR Response (Datasets) Location |
|---|---|---|
| 6/20/2016 | 6/27/2016 | • For Natural History study 901, there were no case report forms used in the collection of the data points; the information was transcribed from the source documents (medical charts).<br>• Study 190-901 is a natural history registry rather than a clinical trial; patients were not required to have clinic visits at specific intervals but as deemed necessary by the investigator for a specific patient. Therefore, intervals between clinic visits vary a lot.<br>\\cdsesub1\evsprod\BLA761052\0003\m1\us\111-information-amendment\1113-efficacy-information-amendment |
| 6/20/2016 | 6/30/2016 | • Patient profiles for 901 subjects based on motor, language and combined motor plus language scores.<br>• Perform the 1:1 matching analysis based on baseline CLN2 score and age±3 months as well as CLN2 score, age±3 months and common genotype. When more than one match occurred the selection was narrowed further by matching on additional variables in the order (1) detailed genome (2) sex (3) country. The matching analyses were all based on rate of decline and the population which excluded one early terminated 201 subject.<br><br>\\cdsesub1\evsprod\BLA761052\0005\m1\us\111-information-amendment\1113-efficacy-information-amendment |
| 6/20/2016 | 7/11/2016 | Provide the raw data in SAS transport format for the cross-sectional Cornell cohort that you planned to include in your initial control data (\\cdsesub1\evsprod\BLA761052\0007\m5\datasets\190-901-supplement\tabulations\legacy).<br>\\cdsesub1\evsprod\BLA761052\0007\m1\us\111-information-amendment\1113-efficacy-information-amendment |
| 6/20/2016 | 7/27/2016 | • Repeat the MMRM analysis for time in months for a 2 point decline (3-2, 2-1) for motor and language subscales separately for the 190-901 population.<br>• Request longitudinal Cornell data and submit it to the BLA by 8/12/2016 (\\cdsesub1\evsprod\BLA761052\0013\m5\datasets\190-901-supplement\tabulations\legacy)<br>• Determining the Evaluable Population, of the 74 patients available in the DEM-CHILD database as of March 2015i, 41 were ultimately included in the evaluable population for this 190-901 supplemental |

| | | |
|---|---|---|
| | | analysis.<br><br>\\cdsesub1\evsprod\BLA761052\0009\m1\us\111-information-amendment\1113-efficacy-information-amendment |
| 8/8/2016 | 8/12/2016 | Provide dates of birth (not just year) for the 12 subjects from study 201 whom you only provided year of birth. Submit a new DM dataset with this information in SAS format. Repeat matching analyses for all subjects whose actual age is different than the imputed age. Due to German law, the applicant imputed 6/30 as day and month of birth for those 12 patients.<br>\\cdsesub1\evsprod\BLA761052\0014\m1\us\111-information-amendment\1113-efficacy-information-amendment<br>Dataset ADSLUP.xpt was submitted to provide detailed genotype information:<br>\\cdsesub1\evsprod\BLA761052\0014\m5\datasets\190-901-supplement\analysis\adam\datasets |
| 8/17/2016 | 8/25/2016 | • Updated Cornell data submitted on 8/12/2016 but will be delayed due to ongoing negotiation.<br>• Provide a new ADCLN2 dataset in SAS format for study 901 Supplement (complete DEM-CHILDS database) with corrected post-diagnosis flags.<br>\\cdsesub1\evsprod\BLA761052\0019\m5\datasets\190-901-supplement\analysis\adam\datasets<br>\\cdsesub1\evsprod\BLA761052\0019\m1\us\111-information-amendment\1113-efficacy-information-amendment |
| 7/20/2016 | 8/29/2016 | • Updated efficacy data including an additional nine (9) months by performing a data-cut in June 2016.<br>• Additional responder analyses (defined as absence of an unreversed 2 point decline) based on change in M-L score over 84 weeks. Also for motor and language score separately (defined as an absence of an unreversed 1 point decline) over 84 weeks. Furthermore, perform sensitivity analyses based on rate of decline (slope)<br>• Updated patient profiles and descriptive analyses.<br>\\cdsesub1\evsprod\BLA761052\0021\m1\us\111-information-amendment\1113-efficacy-information-amendment |
| 8/31/2016 | 9/02/2016 | Submit the updated source (SDTM) and analysis (ADAM) datasets including an additional nine (9) months by performing a data-cut June 3, 2016, in compliance with the SDS guidance.<br>\\cdsesub1\evsprod\BLA761052\0023\m5\datasets<br>\\cdsesub1\evsprod\BLA761052\0023\m1\us\111-information-amendment\1113-efficacy-information-amendment |
| 9/2/2016 | 9/13/2016 | Define file, reviewer's guide and SAS programs submitted<br>\\cdsesub1\evsprod\BLA761052\0026\m5\datasets\190-202\analysis\adam |
| 9/16/2016 | 9/19/2016 | Submit all the datasets, corresponding define files and reviewer's guides as well as SAS programs for additional analyses requested by the FDA (ADMATCH3.xpt and ADMATCH4.xpt)<br>\\cdsesub1\evsprod\BLA761052\0030\m5\datasets\ise\analysis\adam |
| 8/17/2016 | 10/7/2016 | • Repeat the 1:1 matching and many-to-one matching for baseline M-L CLN2 score, genotype and age (within 3 months) by imputing a DOB of 1/1 and a DOB of 12/31 for each of these subjects (so that these 12 subjects will each have 2 imputed DOB)<br>• In addition to the previously requested analyses on your 6/2016 data- |

| | | |
|---|---|---|
| | | cut, perform 1:1 matching for ITT subjects in 201/202 based on baseline CLN2 motor score, genotype, and age within 3 months compared to the analyzable DEM-CHILDs supplement population and perform 1:1 matching for ITT subjects in 201/202 based on baseline CLN2 language score, genotype, and age within 3 months compared to the analyzable DEM-CHILDs supplement population. \\cdsesub1\evsprod\BLA761052\0039\m1\us\111-information-amendment\1113-efficacy-information-amendment |
| 10/21/2016 | 10/26/2016 | • Submit a table and a SAS transportable dataset for the 201/202 German subjects who had an imputed DOB in the original submission of your application with the updated DOB.<br>• Based on the 6/2016 data-cut, perform 1:1 matching for ITT subjects in 201/202 based on baseline CLN2 motor score, genotype, and age (based on revised DOB for German subjects) within 3 months for only motor scores compared to the analyzable DEM-CHILDs supplement population. Perform 1:1 matching for ITT subjects in 201/202 based on baseline CLN2 language score, genotype, and age (based on revised DOB for German subjects) within 3 months for only language scores compared to the analyzable DEM-CHILDs supplement population.<br>(\\cdsesub1\evsprod\BLA761052\0046\m5\datasets)<br>\\cdsesub1\evsprod\BLA761052\0046\m1\us\111-information-amendment\1113-efficacy-information-amendment |
| 10/27/2016 | 11/2/2016 | There were transcription errors for the 901 dataset originally submitted for BLA submission<br>Corrected 901 datasets submitted (SDTM and ADAM); ISE datasets based on corrected 901 (N=42) and updated 202 data for DOB:<br>\\cdsesub1\evsprod\BLA761052\0050\m5\datasets<br>Responder analysis (absence of an unreversed 2 point decline or score of 0 on the Hamburg Motor-Language, Motor or Language CLN2 scale score over 83 weeks) in the 190-201 ITT population matched for age (+/-3 months) and baseline CLN2 score to the corrected DEMCHILD population. Time to event analysis.<br>\\cdsesub1\evsprod\BLA761052\0050\m1\us\111-information-amendment |
| 10/31/2016 | 11/10/2016 | Updated datasets, programs, define files and reviewer's guides used for responding to IR #19 Question 1 (updated ISE analyses) by November 10, 2016 as requested by FDA. The updated ISE (Question 1) and the MMRM analyses (Question 3) will be provided to FDA on November 15th and 18th respectively as previously communicated to FDA.<br>\\cdsesub1\evsprod\BLA761052\0054\m5\datasets\ise\analysis\adam |
| 10/31/2016 | 11/16/2016 | Updated ISE report based on the corrected Study 190-901 data and Study 190-201/202 data (June 3, 2016 data cut)<br>\\cdsesub1\evsprod\BLA761052\0056\m5\53-clin-stud-rep\535-rep-effic-safety-stud\cln2\5353-rep-analys-data-more-one-stud\ise<br>Suppqs.xpt: \\cdsesub1\evsprod\BLA761052\0056\m5\datasets\190-202\tabulations\sdtm<br>\\cdsesub1\evsprod\BLA761052\0056\m1\us\111-information-amendment\1113-efficacy-information-amendment |
| 10/31/2016 | 11/22/2016 | MMRM analysis results submitted |

<table>
<tr><td></td><td></td><td>\\cdsesub1\evsprod\BLA761052\0059\m1\us\111-information-amendment\1113-efficacy-information-amendment</td></tr>
<tr><td>11/16/2016</td><td>11/23/2016</td><td>Datasets, SAS programs and matching analysis results for motor only submitted \\cdsesub1\evsprod\BLA761052\0061\m5\datasets\ise\analysis\adam
<table>
<tr><th>Row Number</th><th>Matching factors (Method #1)</th><th>190-901 Patient Population (n=49 or 42)</th><th>Results of Analyses</th></tr>
<tr><td>3</td><td>M, age within 3 months</td><td>49</td><td>The response rate in 201/202 for the Motor domain, defined as a rate <1 point/48 weeks, is 100% with a 35% rate difference compared to 901 (p=0.0083) (Table 3.1). The updated SAS dataset (admmot2a.xpt) and corresponding SAS program (ad-admmot2a-sas.txt) are located in Module 5.3.5.3 of this submission.</td></tr>
<tr><td>4</td><td>M, genotype and age within 3 months</td><td>42</td><td>The response rate in 201/202 for the Motor domain, defined as a rate <1 point/48 weeks, is 100%, with a 45% rate difference compared to 901 (p=0.0012) (Table 4.1). An updated ISE dataset (admtmotx.xpt) and corresponding SAS program (ad-admmot2x-sas.txt) are located in Module 5.3.5.3 of this submission.</td></tr>
</table>
\\cdsesub1\evsprod\BLA761052\0061\m1\us\111-information-amendment\1113-efficacy-information-amendment</td></tr>
<tr><td>11/16/2016</td><td>11/28/2016</td><td>Datasets, SAS programs and matching analysis results for combined motor plus language only submitted \\cdsesub1\evsprod\BLA761052\0063\m5\datasets\ise\analysis\adam
<table>
<tr><th>Row Number</th><th>Matching factors (Method #1)</th><th>190-901 Patient Population (n=49 or 42)</th><th>Results of Analyses</th></tr>
<tr><td>1</td><td>ML, age within 3 months</td><td>49</td><td>The response rate in 201/202 for the ML score, defined as a rate <2 point/48 weeks, is 100% with a 58% rate difference compared to 901 (p=0.0001) (Table 1). The updated SAS dataset (admtch3a.xpt) and corresponding SAS program (ad-admtch3a-sas.txt) are located in Module 5.3.5.3 of this submission.</td></tr>
<tr><td>2</td><td>ML, genotype and age within 3 months</td><td>42</td><td>The response rate in 201/202 for the ML score, defined as a rate <2 point/48 weeks, is 100%, with a 41% rate difference compared to 901 (p=0.0072) (Table 2). An updated ISE dataset (admtch4x.xpt) and corresponding SAS program (ad-admtch4x-sas.txt) are located in Module 5.3.5.3 of this submission.</td></tr>
</table>
\\cdsesub1\evsprod\BLA761052\0063\m1\us\111-information-amendment\1113-efficacy-information-amendment</td></tr>
<tr><td>11/16/2016</td><td>12/2/2016</td><td>Datasets, SAS programs and matching analysis results for language only submitted \\cdsesub1\evsprod\BLA761052\0065\m5\datasets\ise\analysis\adam</td></tr>
</table>

<table>
<tr><td></td><td></td><td>
<table>
<tr><th>Row Number</th><th>Matching factors (Method #1)</th><th>190-901 Patient Population (n=49 or 42)</th><th>Results of Analyses</th></tr>
<tr><td>5</td><td>L, age within 3 months</td><td>49</td><td>The response rate in 201/202 for the L score, defined as a rate <1 point/48 weeks, is 100% with a 50% rate difference compared to 901 (p=0.0010) (Table 5.1). One 190-201/202 subject had a language score of 0 at the 300 mg baseline and thus did not have an estimable slope. This subject's results are represented as "NA" in Table 5.1.<br><br>The updated SAS dataset (admlng2a.xpt) and corresponding SAS program (ad-admlng2a-sas.txt) are located in Module 5.3.5.3 of this submission.</td></tr>
<tr><td>6</td><td>L, genotype and age within 3 months</td><td>42</td><td>The response rate in 201/202 for the L score, defined as a rate <1 point/48 weeks, is 100%, with a 50% rate difference compared to 901 (p=0.0010) (Table 6.1). One 190-201/202 subject had a language score of 0 at the 300 mg baseline and thus did not have an estimable slope. This subject's results are represented as "NA" in Table 6.1.<br><br>The updated ISE dataset (admtlngx.xpt) and corresponding SAS program (ad-admtlngx-sas.txt) are located in Module 5.3.5.3 of this submission.</td></tr>
</table>
\\cdsesub1\evsprod\BLA761052\0065\m1\us\111-information-amendment\1113-efficacy-information-amendment
</td></tr>
<tr><td>12/23/2016</td><td>12/27/2016</td><td>Additional matching analyses 42/22 (screening and 300mg baseline) for each of Motor, Language and Motor-Language score using different matching criteria for both 48 and 72 weeks. Both McNemar exact and Fisher's exact test need to be used. For missing data, last available score and next observation carried backward (NOCB) should be implemented. In addition, plan for Bayesian approach, proposal for ordinal analysis and duration analysis need to be performed as the Agency requested. Early terminated Study 201 subject 190201-1287-1007 needs to be included in the all the above analyses.<br>\\CDSESUB1\evsprod\BLA761052\0074\m1\us\111-information-amendment\1113-efficacy-information-amendment</td></tr>
<tr><td>12/23/2016</td><td>1/4/2017</td><td>Matching analysis results and time to event analysis results for motor only submitted<br>\\CDSESUB1\evsprod\BLA761052\0075\m1\us\111-information-amendment\1113-efficacy-information-amendment</td></tr>
<tr><td>12/23/2016</td><td>1/10/2017</td><td>Matching analysis results and time to event analysis results for Motor-Language only submitted<br>\\CDSESUB1\evsprod\BLA761052\0076\m1\us\111-information-amendment\1113-efficacy-information-amendment<br>Matching datasets, define file, reviewer's guide and SAS programs submitted<br>\\CDSESUB1\evsprod\BLA761052\0076\m5\datasets\ise\analysis\adam</td></tr>
<tr><td>1/11/2017</td><td>1/12/2017</td><td>More datasets, define file, reviewer's guide and SAS programs submitted<br>\\CDSESUB1\evsprod\BLA761052\0077\m5\datasets</td></tr>
<tr><td>12/23/2016</td><td>1/13/2017</td><td>Matching analysis results and time to event analysis results for Motor-Language only submitted<br>\\CDSESUB1\evsprod\BLA761052\0080\m1\us\111-information-amendment\1113-efficacy-information-amendment<br>Matching datasets, define file, reviewer's guide and SAS programs submitted<br>\\CDSESUB1\evsprod\BLA761052\0080\m5\datasets\ise\analysis\adam</td></tr>
<tr><td>1/19/2017</td><td>1/24/2017</td><td>Four ordinal analysis datasets (42/22), SAS programs, define file and reviewer's guide submitted<br>\\CDSESUB1\evsprod\BLA761052\0084\m5\datasets\analysis-combined\analysis\adam</td></tr>
</table>

| | | |
|---|---|---|
| | | \\CDSESUB1\evsprod\BLA761052\0084\m1\us\111-information-amendment\1113-efficacy-information-amendment |
| 1/19/2017 | 1/25/2017 | Additional analyses (time to event, responder analysis of logistic regression and categorical response) results for motor only based on 72-week data submitted<br>\\CDSESUB1\evsprod\BLA761052\0085\m5\datasets<br>\\CDSESUB1\evsprod\BLA761052\0085\m1\us\111-information-amendment\1113-efficacy-information-amendment |
| 9/1/2016 | 1/31/2017 | CORNELL data submitted<br>\\CDSESUB1\evsprod\BLA761052\0089\m5\datasets\190-901-supplement\tabulations\legacy\datasets |
| 2/1/2017 | 2/6/2017 | One missing dataset submitted (ADDRS.xpt):<br>\\CDSESUB1\evsprod\BLA761052\0090\m5\datasets\190-202\analysis\adam<br>\\CDSESUB1\evsprod\BLA761052\0090\m1\us\111-information-amendment\1113-efficacy-information-amendment<br>Updated Study 201/202 raw data to 96 weeks (November 1, 2016 data cutoff) , define file and reviewer's guide submitted<br>\\CDSESUB1\evsprod\BLA761052\0091\m5\datasets\190-202\tabulations\sdtm\datasets |
| 2/1/2017 | 2/16/2017 | Efficacy analysis results using these 96 week data submitted<br>\\CDSESUB1\evsprod\BLA761052\0096\m1\us\111-information-amendment\1113-efficacy-information-amendment<br>Datasets (42/22), SAS programs, define file and reviewer's guide submitted<br>\\CDSESUB1\evsprod\BLA761052\0096\m5\datasets |
| 2/17/2017 | 2/27/2017 | Data discrepancy issues addressed<br>\\CDSESUB1\evsprod\BLA761052\0104\m1\us\111-information-amendment\1113-efficacy-information-amendment |
| 2/21/2017 | 3/9/2017 | Supplemental SAP submitted<br>\\CDSESUB1\evsprod\BLA761052\0107\m1\us\111-information-amendment\1113-efficacy-information-amendment |
| 3/9/2017 | 3/15/2017 | Response to FDA comments on supplemental SAP<br>\\CDSESUB1\evsprod\BLA761052\0116\m1\us\111-information-amendment\1113-efficacy-information-amendment |
| 3/22/2017 | 3/24/2017 | Part1 of the analysis results submitted<br>\\CDSESUB1\evsprod\BLA761052\0117\m1\us\111-information-amendment\1113-efficacy-information-amendment<br>Datasets, SAS programs, define file and reviewer's guide submitted<br>\\CDSESUB1\evsprod\BLA761052\0117\m5 |
| 3/22/2017 | 3/27/2017 | Part2 of the analysis results submitted<br>\\CDSESUB1\evsprod\BLA761052\0119\m1\us\111-information-amendment\1113-efficacy-information-amendment<br>Datasets, SAS programs, define file and reviewer's guide submitted<br>\\CDSESUB1\evsprod\BLA761052\0119\m5 |
| 4/12/2017 | 4/14/2017 | Matching analysis results (42/24) for motor only submitted<br>\\CDSESUB1\evsprod\BLA761052\0125\m1\us\111-information-amendment\1113-efficacy-information-amendment<br>Datasets, SAS programs, define file and reviewer's guide submitted<br>\\CDSESUB1\evsprod\BLA761052\0125\m5 |

**Table 23. Summary of Univariate Analysis for Time-to-decline Analysis**

| Covariates | Test Method | P-value |
|---|---|---|
| sex | Log-rank | 0.1792 |
| birth year (<=2000 (Y/N)) | Log-rank | 0.0176 |
| screening age | Cox-regression | 0.2226 |
| Initial motor score | Cox-regression | <0.0001 |
| Genotype (0 key mutation (Y/N)) | Log-rank | 0.0233 |
| Genotype (2 key mutations (Y/N)) | Log-rank | 0.7774 |

**Table 24. Summary of AIC Values for Model Selection**

| Model | AIC |
|---|---|
| Trt birthn | 138.971 |
| Trt agescrenbl | 139.063 |
| Trt aval (baseline) | 126.572 |
| Trt birthn agescrenbl | 140.934 |
| Trt birthn aval | 128.302 |
| Trt agescrenbl aval | 128.572 |
| Trt birthn agescrenbl aval | 130.295 |
| Trt GC1 | 130.042 |
| Trt birthn GC1 | 131.886 |
| Trt GC1 agescrenbl | 131.232 |
| Trt GC1 aval | 122.176 |
| Trt birthn GC1 agescrenbl | 133.165 |
| Trt birthn GC1 aval | 124.053 |
| Trt GC1 agescrenbl aval | 124.173 |
| Trt GC1 agescrenbl aval birthn | 126.041 |

Notations
Trt:Study 201/202 and Study 901 (0 and 1)
agescrenbl: screening age
birthn: birth year (<=2000 (Y/N))
GC1: genotype (0 key mutation (Y/N))
aval: screening motor score

---

**This is a representation of an electronic record that was signed electronically and this page is the manifestation of the electronic signature.**

---

/s/

---

LILI GARRARD
04/26/2017

MIN MIN
04/26/2017

SCOTT S KOMO
04/26/2017
I concur

YEH FONG CHEN
04/26/2017

LAURA L JOHNSON
04/26/2017
I concur

# STATISTICAL REVIEW AND EVALUATION
## FILING REVIEW OF AN NDA/BLA

| | |
|---|---|
| **NDA/BLA #:** | BLA 761052 |
| **Related IND #:** | 122472 |
| **Product Name:** | BMN190 (cerliponase alfa) |
| **Indication(s):** | The treatment of patients with CLN2 disease, also known as tripeptidyl peptidase-1 (TPP1) deficiency. |
| **Applicant:** | BioMarin |
| **Dates:** | Stamp Date: 05/27/2016<br>Primary Reviews: 10/27/2016<br>PDUFA Goal date: 01/27/2017 |
| **Review Priority:** | Priority |
| **Biometrics Division:** | DB III |
| **Statistical Reviewer:** | Min Min, Ph.D. |
| **Concurring Reviewers:** | Yeh-Fong Chen, Team Leader, Ph.D.<br>Lili Garrard, Ph.D.<br>Scott Komo, Ph.D. |
| **Medical Division:** | Division of Gastrointestinal and Inborn Error Products (DGIEP) |
| **Clinical Team:** | Medical Officer: Elizabeth Hart, M.D.<br>Medical Team Leader: Laurie Muldowney, M.D. |
| **Project Manager:** | Jenny Doan, |

# 1. Summary of Efficacy/Safety Clinical Trials to be Reviewed

**Table 1: Summary of Trials to be Assessed in the Statistical Review**

| Clinical Studies | | |
|---|---|---|
| 190-201 (Phase 1/2) (Multicenter, Open-label, Dose escalation study)<br><br>Completion of 48 weeks dosing in the stable dose period. | Evaluate safety and tolerability of BMN 190;<br><br>Evaluate effectiveness of BMN 190 by change in ML scale score | **BMN 190 24 patients age 3-15:**<br>Dose Escalation Period: 30 mg, 100 mg, 300 mg every other week<br><br>Stable Dose Period: 300 mg every other week<br><br>ICV infusions |
| 190-202 (Phase 1/2) (Multicenter, Open- | Evaluate long-term safety of BMN 190 | **BMN190 23 patients age 3-15:**<br>300 mg every other week |

| | | |
|---|---|---|
| label study)<br>Up to 240 weeks | Assess change in ML scale score | ICV infusions |

Source: Sponsor's Table 2.7.3.1.4.1 of summary-clin-efficacy.pdf

# 2. Assessment of Protocols and Study Reports

**Table 2: Summary of Information Based Upon Review of the Protocol(s) and the Study Report(s)**

| Content Parameter | Response/Comments |
|---|---|
| Designs utilized are appropriate for the indications requested. | **Yes** |
| Endpoints and methods of analysis are specified in the protocols/statistical analysis plans. | **No, primary efficacy endpoint was not pre-specified** |
| Interim analyses (if present) were pre-specified in the protocol with appropriate adjustments in significance level. DSMB meeting minutes and data are available. | **No** |
| Appropriate details and/or references for novel statistical methodology (if present) are included (e.g., codes for simulations). | **NA** |
| Investigation of effect of missing data and discontinued follow-up on statistical analyses appears to be adequate. | **NA** |

# 3. Electronic Data Assessment

**Table 3: Information Regarding the Data**

| Content Parameter | Response/Comments |
|---|---|
| Dataset location | \\CDSESUB1\evsprod\BLA761052\0000 |
| Were analysis datasets provided? | Yes |
| Dataset structure (e.g., SDTM or ADaM) | SDTM and ADaM |
| Are the define files sufficiently detailed? | Yes |
| List the dataset(s) that contains the primary endpoint(s) | ADSL and ADDRS |
| Are the *analysis datasets* sufficiently structured and defined to permit analysis of the primary endpoint(s) without excess data manipulation? * | No flags for the matching method #1 and #2 |
| Are there any initial concerns about site(s) that could lead to inspection? If so, list the site(s) that you request to be inspected and the rationale. | Roman (similar effect size but larger population than the other site) |
| Safety data are organized to permit analyses across clinical trials in the NDA/BLA. | Only one trial with extension and historical control |

* This might lead to the need for an information request or be a refuse to file issue depending on the ability to review the data.

# 4. Filing Issues

**Table 4: Initial Overview of the NDA/BLA for Refuse-to-file (RTF):**

| Content Parameter | Yes | No | NA | Comments |
|---|---|---|---|---|
| Index is sufficient to locate necessary reports, tables, data, etc. | Yes | | | |
| ISS, ISE, and complete study reports are available (including original protocols, subsequent amendments, etc.) | Yes | | | Since the sponsor only sent the ISE protocol for review, the primary efficacy results will be based on the ISE. |
| Safety and efficacy were investigated for gender, racial, and geriatric subgroups investigated. | | | NA | Very small sample size |
| Data sets are accessible, sufficiently documented, and of sufficient quality (e.g., no meaningful data errors). | Yes | | | Natural history data has missing matching methods #1 and #2. Also, the information for data resources was not provided |
| Application is free from any other deficiency that render the application unreviewable, administratively incomplete, or inconsistent with regulatory requirements | | | | Not sure about this |

**IS THE APPLICATION FILEABLE FROM A STATISTICAL PERSPECTIVE?**
**Yes / No**
**Yes**

# 5. Comments to be Conveyed to the Applicant

## *5.1. Refuse-to-File Issues*

In general, no substantial statistical issues were identified. However, there were many open IRs sent to the Applicant, including IRs for the full evidence dossier.

## *5.2. Information Requests/Review Issues*

- Please provide flags for matching methods #1 and #2.
- Conduct the same analyses for extension trial 202 based on ISE protocol and follow the FDA's recommendations.
- For Study 201/202, provide a list of video file names for all the clinical assessments at all time points that were videotaped.

- The natural history Study 901 serves as the external control for Study 201/202 patients. However, Study 901 used the original Hamburg scale and Study 201/202 used the adapted Hamburg scale. As part of the BLA submission, the Applicant submitted a full evidence dossier to support the validation of the adapted Hamburg scale, and address scale comparability between the original Hamburg scale and the adapted Hamburg scale. Both the validation and scale comparability analyses were conducted using only subsets of the videos from Study 201/202. The adequacy of these analyses will be a review issue.

---

**This is a representation of an electronic record that was signed electronically and this page is the manifestation of the electronic signature.**

---

/s/

---

MIN MIN
07/28/2016

YEH FONG CHEN
07/28/2016