

BLA Clinical Outcome Assessment Review Memorandum

Reviewer names	<p>Naomi Knoble, PhD, Licensed Psychologist Associate Director Division of Clinical Outcome Assessment (DCOA) Office of Drug Evaluation Sciences (ODES) Office of New Drugs (OND) Center for Drug Evaluation and Research (CDER)</p> <p>David Reasner, PhD Director, DCOA, ODES, OND, CDER</p>
Consulted by	<p>Division of Clinical Evaluation General Medicine Branch 1 Office of Clinical Evaluation (OCE) Office of Therapeutic Products (OTP) Center for Biologics Evaluation and Research (CBER)</p>
COA tracking number	C2023261
BLA#	125758
Established name	Atidarsagene autotemcel
(Proposed) trade name	LENMELDY
Applicant	Orchard Therapeutics
Proposed Indication	<p>Treatment of pediatric metachromatic leukodystrophy (MLD): pre-symptomatic late infantile (PSLI) pre-symptomatic early juvenile (PSEJ) early symptomatic early juvenile (ESEJ)</p> <p><input checked="" type="checkbox"/> Rare Disease/Orphan Designation/RMAT <input checked="" type="checkbox"/> Pediatric</p>
Instrument(s) reviewed	<p>Bayley Scale of Infant and Toddler Development Wechsler Preschool and Primary Scale of Intelligence Wechsler Intelligence Scale for Children Wechsler Adult Intelligence Scale</p> <p><input checked="" type="checkbox"/> Performance outcome (PerfO)</p>

Table of Contents

Tables 2

1.	Executive Summary.....	3
1.1.	Review Issues	3
1.2.	Conclusions	4
2.	Brief Regulatory Background	4
3.	Patient Experience Data	4
4.	Efficacy Study Designs and COA-Based Endpoints	5
4.1.	COA-Based Efficacy Endpoints	6
5.	Neurocognitive Test Review	9
5.1.	Neurocognitive Test Descriptions	9
5.2.	Neurocognitive Assessment Frequency.....	11
5.3.	Neurocognitive Assessment Administration and Standardization	12
6.	Neurocognitive COA-based Endpoint Review	13
6.1.	Neurocognitive review from study 201222	13
6.2.	Neurocognitive review from expanded access subject (b) (6)	15
6.3.	Reviewer conclusions on the neurocognitive data review	16
	Appendices.....	16
	Appendix A. Materials Reviewed	17
	Appendix B. Review issues with neurocognitive assessment administration and standardization	20

Tables

Table 1.	Review Issues and Conclusions	3
Table 2.	COA-Based Efficacy Endpoints and Reviewer Comments	7
Table 3.	Neurocognitive Assessment Performance Scores	9
Table 4.	Pre-Symptomatic Early Juvenile Subjects Neurocognitive Data Review.....	13
Table 5.	Early Symptomatic/Early Juvenile Subject Neurocognitive Data Review	13

1. Executive Summary

This review of the Applicant’s neurocognitive data is in response to the Division of Clinical Outcome Assessment (DCOA) consult request by Center for Biologics Evaluation and Research (CBER), Office of Therapeutic Products (OTP), Office of Clinical Evaluation (OCE), Division of Clinical Evaluation General Medicine Branch 1 (GMB1) on August 25, 2023. This review request is related to the cognitive functioning evidence submitted as a COA-based efficacy endpoint for BLA 125758 for atidarsagene autotemcel (OTL-200) a proposed treatment for pediatric metachromatic leukodystrophy (MLD), including pre-symptomatic late infantile (PSLI), pre-symptomatic early juvenile (PSEJ), and early symptomatic early juvenile (ESEJ) of which PSEJ and ESEJ data were the focus of this review.

The sources for this review are: (1) the licensing application; (2) sources external to the application, including publications; and (3) additional materials submitted by the Applicant as requested by the Agency during the review process. See [Appendix A](#) for a listing of materials reviewed, information requests sent, and publications referenced in this review.

1.1. Review Issues

See Table 1 for review issues identified in the Applicant’s submission and reviewer conclusions.

Table 1. Review Issues and Conclusions

	Deficiency	Conclusion
1	The IQ-based eligibility criteria threshold for ESEJ subjects in study 201222 of intelligence quotient (IQ) ≥ 70 resulted in a higher baseline level of cognitive functioning for some and/or all ESEJ subjects than the natural history subjects making natural history subject comparisons infeasible.	The natural history of MLD indicates that early juvenile-onset subjects may present initially with a combination of motor and cognitive symptoms leading to severe motor and cognitive decline over time, including loss of language (Kehrer et al., 2014; Kehrer et al., 2021). Using age-normed standardized scores (i.e., performance standard scores), the trajectory of the treated subjects can be understood in comparison to typically developing children and to their own trajectory. Published evidence from HSCT-treated patients was also used for comparison.
2	The Applicant’s proposed labeling language based on age equivalent scores and their derivatives (i.e., developmental quotient) is not acceptable for regulatory decision-making given the scientific limitations of age equivalent scores.	Performance standard scores convey functioning relative to typically developing same-age peers and are considered acceptable given that the Applicant’s endpoint definition (severe cognitive impairment < 55) has clear clinical meaningfulness and the relatively long-term follow-up evidence available.
3	Issues in the selection and standardized administration of the neurocognitive tests included: (1) fine motor impairments interfered with subject performance on some	Given the clinical meaningfulness of the Applicant’s COA-based endpoint defined as first occurrence of severe cognitive impairment (performance standard score ≤ 55 maintained at all subsequent

subtests and, thereby, confounded scores, (2) non-standardized protocol-driven neurocognitive test selection for subject age and/or ability, (3) non-standardized protocol-driven subtest selection, (4) non-standardized administration when incorporating linguistic interpreters, and (5) no quality review oversight of neurocognitive assessment administration and scoring.	visits) or death from any cause, the importance and relevance of cognitive functioning for MLD patients, and the available long-term data, the COA-based endpoint is interpretable and sufficient for regulatory use despite these administration and standardization (i.e., validity, reliability) review issues.
---	--

1.2. Conclusions

Based on this review, the COA-based endpoint using neurocognitive assessment scores for PSEJ and ESEJ subjects in study 201222 can be considered interpretable and sufficient. Using performance standard scores, subject neurocognitive performance over time demonstrates a positive response for the 3 subjects classified as PSEJ and 3 out of 7 subjects classified as ESEJ indicating that OTL-200 slowed cognitive disease progression (e.g., problem-solving, reasoning). Verbal standard scores were also reviewed. In total, this reviewer concludes that these positive responses for some, but not all, ESEJ subjects reflect a clinically meaningful treatment effect that represents a departure from the known natural history of MLD related to cognitive functioning.

The ESEJ subjects with a positive neurocognitive treatment response had other apparent features of disease progression including physical functioning (e.g., declining GMFC-MLD scores reflecting motor impairment). In contrast, these subjects maintained average or near-average range cognitive functioning compared to typically developing same-age peers despite evidence of MLD-related motor progression. Retention of cognitive functioning was reported to be a meaningful outcome in the Voice of the Patient report (Cure MLD, 2022).

2. Brief Regulatory Background

Atidarsagene autotemcel received FDA's Orphan Drug Designation in March 2018, Rare Pediatric Disease Designation in April 2018, and Regenerative Medicine Advanced Therapy Designation (RMAT) in January 2021. Atidarsagene autotemcel was approved for the treatment of all forms of MLD by the European Medicines Agency (EMA) under the trade name LIBMELDY in 2020.¹ Atidarsagene autotemcel is a gene therapy comprising autologous CD34+ cells, prepared from the patient's own hematopoietic stem cells, transduced with a lentiviral vector that encodes the human arylsulfatase A gene, and is delivered in a one-time IV infusion. The product is manufactured for each individual patient.

3. Patient Experience Data

Patient experience data (PED) were submitted in this file. See the clinical review memo for the PED summary table.

¹ For EMA approval documents, see <https://www.ema.europa.eu/en/medicines/human/EPAR/libmeldy>

Reviewer comments on patient experience data

In addition to the clinician-reported outcome (ClinRO) and performance-based outcome (PerfO) evidence submitted, this reviewer also considered the following sources of PED:

1. Externally-led Patient-Focused Drug Development Voice of the Patient reports:
 - a. Cure MLD (2022). Metachromatic Leukodystrophy (MLD) Voice of the Patient Report, October 21, 2022 and November 18, 2022. Accessed from:
https://www.curemld.com/files/ugd/db3510_93de15438f6b4dd4b2ab87a37c0425d6.pdf
 - b. Cure MLD (2022). Metachromatic Leukodystrophy (MLD) Voice of the Patient: Additional Patient Comments. Accessed from:
<https://mldpfdd.org/voiceofthepatient/#comments>
2. Patient-Centered Research Publications
 - a. Harrington, M., Whalley, D., Twiss, J. et al. Insights into the natural history of metachromatic leukodystrophy from interviews with caregivers. Orphanet J Rare Dis 14, 89 (2019).

These additional PED helped confirm the importance of preserving cognitive functioning for all MLD patients but, in particular, for juvenile patients for whom cognitive symptoms may be an early emerging symptom. In the MLD Voice of the Patient report, caregivers indicated a top concern was decreased communication/responsiveness for MLD patients and the report described patients as “locked in” which is assumed to mean complete dependence on caregivers and no communication abilities, even minimally (e.g., blinking yes/no). Slowing disease progression and increasing responsiveness for patients were reported as valued aspects of a treatment for MLD. These PED further emphasize the substantial unmet treatment need and the importance of neurocognitive preservation as a treatment outcome for patients of all MLD subtypes.

4. Efficacy Study Designs and COA-Based Endpoints

Study 201222 (formerly titled TIGET-MLD), an open-label, non-randomized study of OTL-200 in approximately 30 subjects with MLD. Efficacy endpoint interim analyses were conducted at 3-years with efficacy and safety follow-up analyses conducted at 8-years post-treatment (see protocol 201222 version 13.2, STN 0002). Specific to neurocognitive assessments, an eligibility criterion for ESEJ patients specified an intelligence quotient (IQ) ≥ 70 . No neurocognitive score entry thresholds were specified for other MLD subtypes (see page 76 of 198, protocol 20222 version 13.2).

Study 205756 was a single-arm, open-label, non-randomized study of OTL-200 in approximately 10 subjects with MLD (protocol version 7.2, amendment 7, effective November 18, 2021). Specific to neurocognitive assessment, an eligibility criterion for ESEJ subjects specified an IQ < 85 for exclusion.

Reviewer comments on IQ-based eligibility criteria

Given the relatively shorter follow-up duration in study 205756, these data were not the focus of this review.

From a clinical trial design perspective, the IQ threshold criterion should have been applied to PSEJ subjects as well as ESEJ subjects. While PSEJ subjects by clinical definition are not yet exhibiting features of MLD, a range of cognitive functioning above and below the average range would be expected for this population. While this did not impact the current review, it is noted for future MLD development programs.

By specifying the ESEJ subject eligibility IQ threshold at ≥ 70 in study 201222, all enrolled ESEJ subjects were not at the lowest end of the cognitive functioning range (i.e., severely impaired) at baseline. While these thresholds enable the assessment of maintenance and deterioration of cognitive function, they also likely resulted in some and/or all early juvenile subjects having a higher baseline level of cognitive functioning than natural history subjects. As a point of comparison, evidence from juvenile MLD subjects who received hematopoietic stem cell transplantation indicates that average or near-average range neurocognitive functioning was characteristic of PSEJ and ESEJ subjects at the time of transplantation (e.g., Beschle et al., 2020; Groeschel et al., 2016). Following from this published evidence, the baseline neurocognitive functioning of ESEJ subjects eligible for protocol 201222 is reflective of PSEJ and ESEJ patients in treatment, but does not encompass the full range of neurocognitive functioning as the thresholds exclude severe impairment.

Natural history studies of MLD indicate that some subjects may present with motor and cognitive symptoms and experience cognitive decline over time leading to loss of language (Kehrer et al., 2014; 2021). Across many rare diseases, natural history subjects tend to be older in chronological age and/or may have a more severe disease presentation than subjects enrolled in clinical treatment trials. It may be difficult to identify ESEJ subjects early in the disease course before the emergence of cognitive and/or motor impairments, which may help explain the relatively severe presentation of the MLD natural history population. These differences can make subject comparisons difficult or infeasible, and may inflate apparent differences between treated subjects and natural history controls when the entry criteria for clinical treatment trials exclude more severe subjects.

However, by using population-normed standardized scores (i.e., performance standard scores, normalized to Mean=100 and Standard Deviation=15) from the neurocognitive assessments administered in study 201222, the trajectory of treated subjects can be understood in comparison to a population of typically developing children and to the subject's prior performance.

4.1. COA-Based Efficacy Endpoints

The Applicant's COA-based efficacy endpoint based on neurocognitive assessments used the *performance standard scores*² from these assessments (see study 201222

² The term *performance standard score* was used in the Applicant's submission materials as a term to summarize the following scores from the selected neurocognitive assessments: Bayley-3 Cognitive scale; WPPSI-III Perceptual Organization, WPPSI-IV Visual Spatial Index and Fluid Reasoning Index, WISC-III and -IV Performance Index and Performance Index Quotient, and WAIS-IV Perceptual Reasoning Index.

statistical analysis plan (SAP) version 4.0, pages 66-67 of 100). See Table 2 for COA-based efficacy endpoints and reviewer comments.

Table 2. COA-Based Efficacy Endpoints and Reviewer Comments

Primary	COA	Proposed Labeling (abbreviated)	Reviewer Comments
Severe motor impairment-free survival in the interval from birth to the earlier of either loss of locomotion and sitting without support as measured by the Gross Motor Function Classification MLD (GMFC-MLD) score of level ≥ 5 or death from any cause	GMFC-MLD	Treatment with LENMELDY significantly reduced the risk of severe motor impairment or death in patients with PSLI compared with untreated LI natural history patients.	See clinical outcome assessment review focused on the GMFC-MLD by Dr. Swett.
Key Secondary	COA	Proposed Labeling	Reviewer Comments
Proportion of subjects who experienced severe motor impairment (defined by a GMFC-MLD level ≥ 5) or death, evaluated at 2-years post treatment with OTL-200 for treated subjects, and based on assessments made at matching ages for natural history subjects	GMFC-MLD	A secondary endpoint of the integrated efficacy analysis was motor function, as assessed by the proportion of patients who had experienced severe motor impairment or death by 2-years and 5-years post treatment with LENMELDY. At 2-years post-treatment, a smaller proportion of PSLI patients experienced severe motor impairment or death as compared to age matched natural history subjects.	See primary efficacy endpoint comments.
Additional Secondary	COA	Proposed Labeling	Reviewer Comments
Motor impairment-free survival, defined as the interval from birth to the earlier of loss of ability to walk (GMFC-MLD level ≥ 3 , confirmed at all subsequent visits) or death from any cause	GMFC-MLD	See review by Dr. Swett	See primary efficacy endpoint comments.
Additional Secondary	COA	Proposed Labeling	Reviewer Comments
Performance age-equivalents over time	Neurocognitive tests	Cognitive Performance Age Equivalent: An additional endpoint of the	Age-equivalent scores are not an acceptable score metric to support

<p>post-treatment with OTL-200</p>		<p>integrated efficacy analysis was cognitive function, as assessed by performance age-equivalents over time post-treatment” with claims specific to PSLI and ESEJ patients with associated plots.</p>	<p>regulatory decision making or labeling.</p> <p>Scores resulting from the performance domain of neurocognitive assessments may be acceptable for use in MLD subjects.</p>
<p>Performance standard scores/development quotient (performance) over time post treatment with OTL-200</p>	<p>Neurocognitive tests</p>	<p>No labeling language proposed.</p>	<p>Standardized scores are an acceptable score metric and are suitable to characterize the range of cognitive functioning seen in the MLD population.</p>
<p>Confirmed severe cognitive impairment-free survival, defined as the interval between birth and the first occurrence of severe cognitive impairment (performance standard score ≥ 55 maintained at all subsequent visits) or death from any cause</p>	<p>Neurocognitive tests</p>	<p>No labeling language proposed.</p>	<p>This COA-based efficacy endpoint is interpretable despite validity and reliability issues, and the proposed threshold for severe cognitive impairment can be evaluated with a sensitivity analysis.</p>

5. Neurocognitive Test Review

5.1. Neurocognitive Test Descriptions

See Table 3 for the neurocognitive assessments administered in study 201222. See protocol 201222 version 13.2 section 8.7.3.2 for a listing of neurocognitive tests. The Applicant defined the following categories of cognitive functioning as: normal (performance standard score ≥ 85), mild impairment (performance standard score < 85 and ≥ 70), moderate impairment (performance standard score < 70 and ≥ 55), and severe impairment (performance standard score < 55). See Table 3 for an abbreviated summary of neurocognitive assessments used in study 201222 to derive the performance standard scores.

Table 3. Neurocognitive Assessment Performance Scores

Assessment	Age Range	Scales/Subtests
Bayley Scale of Infant and Toddler Development, 3 rd edition (Bayley-3)	1- to 42-months	Cognitive scale, 91-items Note: The Bayley cognitive scale is a developmental assessment (e.g., sensorimotor development, exploration of objects) and is not an intelligence test.
Wechsler Preschool and Primary Scale of Intelligence, 3 rd edition (WPPSI-III); published in 2004	2.6- to 7.3-years	2.6- to 3.11-years: Block Design Object Assembly 4- to 7.3-years: Block Design Matrix Reasoning Picture Concepts (Picture Completion) (Object Assembly) Note: The Block Design subtest tasks are identical across both age versions of the WPPSI-III. Picture Completion and Object Assembly are supplemental subtests for the 4- to 7.3-year assessment.
Wechsler Scale of Intelligence for Children, 3 rd edition (WISC-III)	6- to 16-years	Perceptual Organization Picture Completion Picture Arrangement Block Design Object Assembly
Wechsler Scale of Intelligence for Children, 4 th edition (WISC-IV)	6- to 16-years	Perceptual Reasoning Index Block Design Picture Concepts Matrix Reasoning (Picture Completion) Note: Picture Completion is a supplemental subtest.

Wechsler Adult Intelligence Scale, 4 th edition (WAIS-IV)	16- to 90-years	<p><i>Perceptual Reasoning Index</i> Block Design Matrix Reasoning Visual Puzzles <i>(Figure Weights)</i> <i>(Picture Completion)</i></p> <p>Note: Figure Weights and Picture Completion are supplemental subtests.</p>
--	-----------------	---

Reviewer comments on neurocognitive assessments and scores

The selected neurocognitive assessments are commonly used in clinical practice; however, there are limitations to these assessments for children with fine motor impairments, see section 5.3 reviewer comments on standardized administration.

Score thresholds

The Applicant’s categories of cognitive functioning align with broadly accepted categories. However, it is important to note that scores from neurocognitive assessments vary in precision and contain measurement error. Particular thresholds will not necessarily provide meaningful clinical information about a subject’s functioning (Burack et al., 2021). For example, a subject with a performance standard score of 60 and another with a score of 54 may function similarly in their daily lives, yet the Applicant’s cutoff score of 55 suggests different conclusions about their cognitive functioning. A sensitivity analysis was conducted by the Agency to evaluate the impact of varying thresholds on the Applicant’s results. In the context of this review, neurocognitive assessment scores are considered within the totality of evidence presented for MLD subjects.

Selection of standard scores

When selecting scores for use in COA-based endpoints to support regulatory decision-making, there is no ideal or perfect score option. The Applicant’s proposal to use age equivalent (AEQ) scores and their derivatives (i.e., developmental quotient) may have value in communicating results of a child’s neurocognitive test performance; however, they have several conceptual, practical, and psychometric limitations which make them unsuitable to support regulatory decision-making. Limitations include, but are not limited to, the following:

1. AEQ scores mis-represent the range of average pediatric functioning. Specifically, AEQs are created by taking the raw score that corresponds to the mean of a given chronological age group (e.g., 1-month interval groups, 3-month intervals, 1-year intervals). This approach artificially narrows and misrepresents the range of average functioning which can lead to under- or over-estimation of functioning.
2. AEQ scores do not represent the concept of cognitive functioning, but the “mental age” of a subject. "Mental age" is not the same as a measure of cognitive abilities.

3. This reviewer is not aware of sufficient validity evidence to consider AEQ scores fit-for-purpose in the context of MLD.

Performance standard scores (Mean=100, Standard Deviation=15) are age-normed to specific chronological age groups and are intended to convey the subject's functioning relative to the distribution of typically developing same-age peers. The following populations were used to derive standard scores:

- Bayley-3: N=1,700 children ages 16-days to 43-months 15-days with sociodemographic characteristics aligned with the 2000 US Census
- WPPSI-III: N=1,700 children ages 2-years 6-months to 7-years 3-months with sociodemographic characteristics aligned with the 2000 US Census
- WPPSI-IV: N=1,700 children ages 2-years 6-months to 7-years 3-months
- WISC-III: N=2,200 children ages 6- to 16-years with sociodemographic characteristics aligned with the 1988 US Census
- WISC-IV: N=2,200 children ages 6- to 16-years with sociodemographic characteristics aligned with the 2000 US Census

Standard scores also have limitations in the context of neurodegenerative conditions, including: (1) known floor effects, e.g., scores do not reflect gains made by children in the lower range of functioning; (2) may obscure skill retention, e.g., a subject may be experiencing cognitive stabilization while their standard scores reflect a decline; and (3) difficulties with interpretation given that scores are not necessarily precise or intended to be sensitive to small changes over time. Nonetheless, the use of performance standard scores is reasonable in the current context of use given the Applicant's endpoint definition of severe cognitive impairment (performance standard score ≤ 55), lack of natural history subjects with comparable functioning at baseline, and the findings from sensitivity evaluations conducted by the Agency to evaluate alternative outcomes.

5.2. Neurocognitive Assessment Frequency

In study 201222, neurocognitive assessments were conducted at screening, baseline, and every 6-months through month 36 (see protocol 201222 version 13.2, Table 7 efficacy assessment schedule, page 86 of 198), and every 6-months through month 96 (see protocol 201222 version 13.2, Table 8 long-term assessment schedule year 3.5-year 8, page 89 of 198).

Reviewer comments on the frequency of neurocognitive assessment

The 6-month frequency of neurocognitive assessment in study 201222 appears reasonable for a longitudinal clinical trial in MLD. In the context of clinical trials, two neurocognitive data points per year allows for closer monitoring of disease progression, may facilitate the detection of safety events, and facilitates interpretation for scores that are vulnerable to measurement error. Learning effects (also called practice effects) are likely to occur for all subjects in the study, with and without cognitive impairments, which may contribute to an apparent inflation of neurocognitive test scores. Scores on the

subtests comprising the performance standard score typically show the greatest practice effects,³ suggesting that subject scores may increase because children have learned the test. In MLD, these potentially inflated scores due to learning effects would not be expected to overcome a decline in scores associated with severe cognitive decline as MLD progresses. Further, the COA-based efficacy endpoint, defined as a cutoff score of severe impairment, does not rely on small changes in continuous scores.

5.3. Neurocognitive Assessment Administration and Standardization

There were no details in the protocol that established standardized procedures for the neurocognitive assessments. The Applicant clarified in a response (received September 28, 2023, STN 0006) to the clinical information request #1 sent September 22, 2023, that there were no standardized test switching criteria for study 201222, rather, that the judgement of the test administrators informed by guidance in the test administration manuals was used to select assessments.

Reviewer comments on administration and standardization

Only neurocognitive data from study 201222 was reviewed as there was insufficient long-term follow-up for study 205756 at the time of the BLA application.

The following review issues were identified in the selection and standardized administration of the neurocognitive tests: (1) fine motor impairments confounded some scores, (2) non-standardized test selection for subject age and/or ability, (3) non-standardized subtest selection, (4) non-standardized administration when using linguistic interpreters, (5) no quality review oversight of neurocognitive assessment administration and scoring. See [Appendix B](#) for discussion of these points.

These administration and standardization issues create scientific uncertainty in the interpretation of neurocognitive scores. Given the Applicant's COA-based endpoint defined as first occurrence of severe cognitive impairment (performance standard score ≤ 55 maintained at all subsequent visits) or death from any cause, the significant unmet treatment need in MLD, the importance of cognitive functioning for MLD subjects, and the long-term nature of the performance and verbal standard score data available for review, the neurocognitive scores are nonetheless considered interpretable and sufficient for regulatory review and evaluation.

In further support, the sensitivity analyses conducted by the biostatistical reviewers indicate that using a cutoff threshold of 50, 60, and 70 does not substantively change the review conclusions based on the neurocognitive scores.

³ WPPSI-III administration and scoring manual, page 13; WISC-III administration and scoring manual, chapter 1, section applications of the WISC-III (PDF page 14 of 282); WISC-IV administration and scoring manual, page 10.

6. Neurocognitive COA-based Endpoint Review

6.1. Neurocognitive review from study 201222

See Tables 4 and 5 for a summary of PSEJ and ESEJ reviewer comments. The use of “broadly average range” refers to standard scores ≥ 85 and the use of “near average range” refers to standard scores between 80 and < 85 . Range definitions used by the Applicant are used below: normal (performance standard score ≥ 85), mild impairment (performance standard score < 85 and ≥ 70), moderate impairment (performance standard score < 70 and ≥ 55), and severe impairment (performance standard score < 55).

Table 4. Pre-Symptomatic Early Juvenile Subjects Neurocognitive Data Review

ID	Treatment Response	Reviewer Comments
(b) (6)*	Apparent positive treatment response	The subject was evaluated with the Bayley-3 (baseline to Month 18; age 1.5- to 3.1-years), WPPSI-IV (Months 24 to 54; age 3.6- to 6.1-years), and WISC-IV (Months 60 to 96; age 6.9- to 9.9-years). Scores across all assessments appeared stable within the broadly average range for both performance and verbal standard scores.
(b) (6)**	Apparent positive treatment response	The subject was evaluated with the WPPSI-IV (baseline to Month 12; age 5.5- to 6.5-years) and WISC-IV (Months 18 to 72, Month 108; age 7- to 14.9-years). Scores across all assessments were in the broadly average range for both performance and verbal standard scores.
(b) (6)	Apparent positive treatment response	The subject was evaluated with the WPPSI-IV (screening to Month 18; age 3.9- to 5.6-years) and WISC-IV (Months 24 to 96; age 6.1- to 12-years). Scores across all assessments were in the broadly average range for both performance and verbal standard scores. Of note, at Month 72 (6-years post-treatment; age 10.4-years) cognitive functioning was assessed at a different testing site due to the COVID-19 pandemic and resulted in the subject's lowest scores achieved in the study, which were still in the average range.
* (b) (6) was reclassified from PSEJ to PSLI during the review.		
** (b) (6) was reclassified from ESEJ to PSEJ during the review.		

Table 5. Early Symptomatic/Early Juvenile Subject Neurocognitive Data Review

ID	Treatment Response	Reviewer Comments
(b) (6)	Apparent treatment failure for cognitive functioning	The subject was administered the Bayley-3 (baseline to Month 6; age 3.2- to 3.7-years), WPPSI-IV (Months 12 to 36; age 4.2- to 6.2-years), WISC-IV (Month 48; age 7.3-years), transitioned back to the Bayley-3 (Months 48 to 72; age 7.3- to 9.4-years) for out of chronological age range assessment through month 72. The subject demonstrated severe cognitive decline.

		<p>Assessment with the WPPSI-IV from Months 12 to 24 (age 4.2- to 5.7-years) indicated broadly average range functioning for both verbal and performance standard scores. Notably, at Month 30 (age 5.7-years) the subject's performance on these two indices diverged where the verbal standard score was recorded in the moderate impairment range and the performance standard score remained in the average range. At Month 36 (age 6.2-years) the verbal standard score was in the severe impairment range and the performance standard score remained in the average range.</p> <p>Approximately one year later when the subject was assessed with the WISC-IV at Month 48 (age 7.3-years), both verbal and performance standard scores were within or near the severe impairment range. Concurrent assessment at Month 48 with the Bayley-3 indicated the subject could correctly match items by size, correctly identify some familiar pictures, and identify simple patterns, but could not sort items by color or count three blocks. Subsequent assessments with Bayley-3 at Months 60 and 72 (age 8.4- and 9.4-years) indicated that the subject did not retain the skills as assessed at Month 48 and demonstrated skill loss.</p> <p>The subject's cognitive functioning apparently progressed more rapidly than motor functioning. At Month 48 when both verbal and performance standard scores were in the severe impairment range, the subject's GMFC-MLD functioning indicated standing was still possible.</p>
(b) (6)	Apparent positive treatment response for cognitive functioning only, not motor	<p>The subject was administered the WISC-III (screening) and WISC-IV (baseline to Month 66, unscheduled visit at approximately Month 96; age 7.1- to 15.5-years). Scores across all assessments were in the broadly average range on performance standard scores. The subject performed in the broadly average range for verbal standard scores from baseline through Month 66, and in the near average range at approximately Month 96.</p> <p>When GMFC-MLD scores showed progression to only trunk control at Month 36 (age 10.2-years), the subject's cognitive functioning remained in the average range. When GMFC-MLD scores showed only head control remained at the last assessment (age 15.5-years), the subject sustained near average range cognitive functioning.</p>
(b) (6)	Apparent positive treatment response for cognitive functioning	<p>The subject was administered the WISC-IV (screening to Month 48; age 11.4- to 15.8-years) and WAIS-IV (Months 54 to 96; age 16.3- to 20-years). Scores across all assessments were in the broadly average range for performance and verbal standard scores in the presence of motor decline as demonstrated by GMFC-MLD scores that declined from standing to pull-to-standing at age 11-years .</p>

(b) (6)	Apparent positive treatment response for cognitive functioning	The subject was administered the WISC-IV (screening to Month 96; age 6.8- to 15.1-years). Scores across nearly all assessments (except screening) were in the broadly average range for performance and verbal standard scores; however, both performance and verbal scores fell in the range of mild cognitive impairment at Month 96. GMFC-MLD scores declined to pull-to-standing by the last assessment.
(b) (6)	Apparent treatment failure	Death of subject at age 7-years. The subject was administered WPPSI-IV at screening and baseline (age 5.6- and 5.7-years) and performed in the average range for verbal and near average range for performance standard scores.
(b) (6)	Apparent treatment failure	The subject was administered the WPPSI-IV (screening to Month 6; age 5.1- to 6-years) and WISC-IV (Months 12 to 24; age 6.5- to 7.4-years) with average range functioning for both verbal and performance standard scores. An adverse event of cognitive disorder at age 9.6-years was recorded by the principal investigator and no further cognitive assessment data were available. Motor decline from walking to standing occurred at approximately age 9-years. The subject's cognitive functioning apparently progressed more rapidly than motor functioning.
(b) (6)	Apparent treatment failure	Death of subject at age 6.6-years. The subject was administered WPPSI-IV at screening and baseline (age 5.8- and 5.9-years) and performed in the average range for verbal and performance standard scores.

6.2. Neurocognitive review from expanded access subject (b) (6)

Neurocognitive assessments from subject (b) (6) from an expanded access program were reviewed as additional supportive evidence for study 201222. The subject was administered the WISC-III (baseline, Month 6; age 7.5- and 8.4-years), WISC-IV (Months 12 to 72; age 8- to 14-years); and WAIS-IV (Month 120; age 17.7-years).

Reviewer comments on MLD-C02 neurocognitive assessments

Subject (b) (6) had near and/or broadly average range performance standard scores through month 72 (age 14-years) and mild impairment range scores at month 120 (age 17-years). Notably the Block Design subtest, which requires fine motor abilities, was consistently the lowest score of the three subtests administered to the subject. The clinical study report indicated the subject was administered the WISC-V, a substantially updated and revised version of the WISC-IV, at age 16-years with a corresponding performance standard score below 70 which qualified as an adverse event of cognitive disturbance. The subject was administered the WAIS-IV at month 120 with a corresponding performance standard score of 77, indicating mild cognitive impairment.

The subject's overall neurocognitive functioning at age 17-years indicates that they retained independent problem-solving abilities and the ability to verbally convey their

perspective. In the presence of mild cognitive impairment, motor decline and other indicators of neurological disease progression, this subject's retention of cognitive functioning skills appears clinical meaningful.

6.3. Reviewer conclusions on the neurocognitive data review

Of the 3 subjects from study 201222 categorized as PSEJ for whom neurocognitive evidence could be evaluated, there was an apparent positive treatment response for all 3 subjects based on the neurocognitive data. Of the 7 subjects from study 201222 categorized as ESEJ for whom neurocognitive evidence could be evaluated, there was an apparent positive treatment response based on neurocognitive data for 3 of 7 subjects.

All 7 ESEJ subjects demonstrated MLD-related impacts including motor impairments (e.g., declining GMFC-MLD scores, fine motor impairments) and other disease features. However, the 3 subjects with an apparent positive treatment response sustained average or near-average range cognitive functioning compared to typically developing same-age peers despite evidence of other features of MLD progression.

While the available natural history patients could not provide a reasonable comparator given baseline differences in cognitive functioning, evaluation of cognitive evidence from HSCT-treated subjects indicates that ESEJ subjects declined by about 20 standard score points or more in cognitive functioning by approximately 60-months (Beschle et al., 2020; Groeschel et al., 2016). Beschle et al. (2020) observed, "It is important to note that patients who suffered from motor deterioration also suffered from cognitive deterioration. Therefore, if deterioration occurred in the first 12 to 18 months after HSCT, both motor and cognitive function declined" (pg. 4 of 9). This HSCT evidence suggests that cognitive and motor decline would be expected to be observed among ESEJ subjects and cognitive functioning, in particular, would be expected to decline to within the mild or more severe cognitive impairment range for most subjects.

Early juvenile MLD is characterized by variability in the rate of decline between and within each subject; however, these six subjects treated with OTL-200 in study 201222 demonstrate unexpected stability at or near the average range of average cognitive functioning over several years. This suggests an interpretable and sufficient treatment effect despite measurement uncertainties. This reviewer concludes that OTL-200 meaningfully preserves cognitive functioning for some, but not all, ESEJ subjects.

Appendices

[Appendix A](#). Materials Reviewed

[Appendix B](#). Review issues with neurocognitive assessment administration and standardization

Appendix A. Materials Reviewed

Document	STN	Date Received
BLA 125758 2.7.3 Summary of clinical efficacy 5.3.5.1 Analysis Datasets (ADaM) 201222 statistical analysis plan (SAP) version 4.0, effective date 10Apr2022 5.3.5.3 ISE treated subjects efficacy narratives 16.1.1 201222 protocol version 13.2, effective date 18Nov2021	0002	06Jul2023
BLA 125758 1.11.3 response to clinical IR#1 dated 22Sep2023	0006	28Sep2023
BLA 125758 1.11.3 response to clinical IR dated 22Sep2023, question 4b	0007	11Oct2023
BLA 125758 1.11.3 Orchard communication to FDA on define files in BLA data package dated 29Sep2023	0009	18Oct2023
BLA 125758 1.11.3 response to Clinical IR#4 dated 05Dec2023 regarding neurocognitive assessment selection and administration 5.3.5.1 201222 table of neurocognitive assessments administered, subtests, raw scores, scaled and composite scores	0020	13Dec2023
BLA 125758 1.11.3 response to clinical IR#4, question 2, part 1 5.3.5.1 201222 neurocognitive data – Bayley; 201222 neurocognitive data – WPPSI-III	0021	15Dec2023
BLA 125758 1.11.3 response to clinical IR#4, question 2, part 2; appendix 1 English versions of neuropsychological test forms; appendix 2 scoring manuals 5.3.5.1 neurocognitive data – WPPSI-III additional data; WPPSI-III; WISC-IV; clinical summaries WISC-IV; WISC-III MLD13	0026	22Dec2023
BLA 125758 1.11.3 response to clinical IR#5, part 1; appendix 1 WAIS-IV test form; appendix 2 WAIS-IV administration and scoring manual 5.3.5.1 201222 neurocognitive data WISC-IV MLD12 108 months; MLD17 96 months; MLD14 5.3.5.4 207394 neurocognitive data WISC-III, WISC-IV, WAIS-IV MLDCO2	0027	09Jan2024
BLA 125758 1.11.3 response to clinical IR#6 dated 19Jan2024, part 1	0034	24Jan2024
BLA 125758	0036	25Jan2024

1.11.3 response to clinical IR#6 dated 19Jan2024, part 2;
appendix 1 panel plots of performance standard score and
brain MRI scores

Publications

Beschle J, Döring M, Kehrer C, et al., Groeschel S. Early clinical course after hematopoietic stem cell transplantation in children with juvenile metachromatic leukodystrophy. *Mol Cell Pediatr.* 2020 Sep 3;7(1):12.

Boucher, A.A., Miller, W., Shanley, R. et al. Long-term outcomes after allogeneic hematopoietic stem cell transplantation for metachromatic leukodystrophy: the largest single-institution cohort report. *Orphanet J Rare Dis* 10, 94 (2015).

Burack JA, Evans DW, Russo N, Napoleon JS, Goldman KJ, Iarocci G. Developmental Perspectives on the Study of Persons with Intellectual Disability. *Annu Rev Clin Psychol.* 2021 May 7;17:339-363.

Harrington, M., Whalley, D., Twiss, J. et al. Insights into the natural history of metachromatic leukodystrophy from interviews with caregivers. *Orphanet J Rare Dis* 14, 89 (2019).

Cure MLD (2022). Metachromatic Leukodystrophy (MLD) Voice of the Patient Report, October 21, 2022 and November 18, 2022. Accessed from:
https://www.curemld.com/files/ugd/db3510_93de15438f6b4dd4b2ab87a37c0425d6.pdf

Cure MLD (2022). Metachromatic Leukodystrophy (MLD) Voice of the Patient: Additional Patient Comments. Access from: <https://mldpfd.org/voiceofthepatient/#comments>

Fumagalli F, Zambon AA, Rancoita PMV et al., Sessa M. Metachromatic leukodystrophy: A single-center longitudinal study of 45 patients. *J Inherit Metab Dis.* 2021 Sep;44(5):1151-1164.

Groeschel S, Kühl JS, Bley AE, et al. Long-term outcome of allogeneic hematopoietic stem cell transplantation in juvenile metachromatic leukodystrophy compared with nontransplanted control patients. *JAMA Neurol.* Published online July 11, 2016. eTable supplement.

Kehrer C, Blumenstock G, Gieselmann V, Krägeloh-Mann I; GERMAN LEUKONET. The natural course of gross motor deterioration in metachromatic leukodystrophy. *Dev Med Child Neurol.* 2011 Sep;53(9):850-855.

Kehrer C, Elgün S, Raabe C, et al., Groeschel S. Association of Age at Onset and First Symptoms With Disease Progression in Patients With Metachromatic Leukodystrophy. *Neurology.* 2021 Jan 12;96(2):e255-e266.

Kehrer C, Groeschel S, Kustermann-Kuhn B, et al., Krägeloh-Mann I; German LEUKONET. Language and cognition in children with metachromatic leukodystrophy: onset and natural course in a nationwide cohort. *Orphanet J Rare Dis.* 2014 Feb 5;9:18.

Martin, S., Harris, N. & Romanus, D. Evaluating meaningful changes in physical functioning and cognitive declines in metachromatic leukodystrophy: a caregiver interview study. *J Patient Rep Outcomes* 7, 70 (2023).

Morton, G., Thomas, S., Roberts, P. et al. The importance of early diagnosis and views on newborn screening in metachromatic leukodystrophy: results of a Caregiver Survey in the UK and Republic of Ireland. *Orphanet J Rare Dis* 17, 403 (2022).

Sanchez et al. Common rater errors for assessment in pediatric rare and orphan disease trials. International Society for CNS Clinical Trials and Methodology (ISCTM) 2018 Autumn Conference poster presentation.

https://isctm.org/public_access/Autumn2018/PDFs/Sanchez-Poster.pdf

[WPPSI-III Administration and Scoring Manual, 2002. NCS Pearson, Inc.](#)

[WISC-III Administration and Scoring Manual, normative data copyright 1991. The Psychological Corporation.](#)

[WISC-IV Administration and Scoring Manual, 2003. NCS Pearson, Inc.](#)

[WISC-IV Technical and Interpretative Manual, 2003. NCS Pearson, Inc.](#)

Appendix B. Review issues with neurocognitive assessment administration and standardization

The following review issues were identified in the selection and standardized administration of the neurocognitive tests: (1) fine motor impairments confounding some subtests, (2) non-standardized test selection for subject age and/or ability, (3) non-standardized subtest selection, 4) non-standardized administration when using linguistic interpreters, and (5) no quality review oversight of neurocognitive assessment administration and scoring.

1. Fine motor impairments confounding some subtests

Some of the items of the Bayley-3 Cognitive scale and some of the subtests (e.g., Block Design) comprising the Wechsler performance score composite from either the WPPSI or WISC (all editions) would not be recommended for MLD subjects given that fine motor control is needed for subtest completion. When a subject has fine motor impairments the resulting scores may not reflect a subject's cognitive abilities, but instead reflect their challenges using their hands due to the disease progression. It is possible to have a protocol-driven solution that all subjects are administered only subtests that do not require hand control.

The Applicant did not standardize the administration to account for fine motor impairments. In a response to the clinical IR #4, the Applicant cited the neurocognitive assessment manuals and indicated it was considered acceptable to substitute a subtest as directed by the clinician administrator. While the approach may be commonly taken in clinical practice, in the context of a clinical trial to support regulatory decision-making this creates non-standardized administration and scores that may not be comparable.

2. Non-standardized test selection for subject age and/or ability

When generating evidence to support regulatory decision-making, there should be standardized, protocol-specified test selection criteria that clearly describe how to select a test for a subject. Standardized test selection criteria are needed in three general instances. First, when there is an overlap in the age of administration of available tests, it may not be clear which test to administer based on the subject's chronological age. For example, the Bayley-3 can be administered up to age 42-months (3.5-years) and the WPPSI-III can be administered beginning at age 2.6-years. Second, some subjects with cognitive impairments may benefit from the administration of a test that is simpler, but that is not designed for their chronological age. While it is acceptable to administer a simpler test to a subject with cognitive impairments, test selection criteria need to be standardized to clarify how this selection is determined. Third, standardized test selection criteria can help minimize errors in the judgement of the test administrator, which may occur, and ensure that resulting scores are comparable. Standardized, protocol-specified test selection criteria typically address these circumstances.

There were no details in the protocol that established standardized procedures for these or other circumstances. The Applicant clarified in a response (received September 28,

2023, STN 0006) to the clinical information request #1 sent September 22, 2023, that there were no standardized test switching criteria for study 201222, but the judgement of the test administrators and guidance in the test administration manuals were used to select assessments. Clinician judgement may not be standardized across the three administrators involved in this study, and the test administration manuals were not written for the level of standardization necessary for clinical trials.

3. Non-standardized subtest selection

It was not clearly described in the protocol or SAP which subtests were administered to derive the resulting performance scores. For example, the subtests of the WPPSI-III and WPPSI-IV were specified in the SAP, but not the WISC-III and WISC-IV. However, it was evident in the subject narratives (ISE-treated subjects efficacy narratives, STN 0002) provided by the Applicant that not all subjects were administered the same subtests to derive the performance scores. For example, in study 201222 subject MLD13 had a GMFC-MLD score that indicated severe motor impairments (i.e., only head control) with a performance standard score within the normal range. If all standardized subtests for the performance standard score were administered, it is unclear how a subject with only head control could use their hands to complete the Block Design subtest. The ADIQCOMM datafile did not provide further clarification. The SAP documented subtests from which performance scores were derived for the Bayley-II, Bayley-III, WPPSI-III, and WPPSI-IV, but not the WISC-III or WISC-IV.

In a response to clinical IR #4 sent December 5, 2023 (received December 13, 2023, STN 0020), the Applicant clarified that subtest substitutions were made in study 201222 and provided a full datafile of all subtests administered for all subjects at all timepoints. The Applicant indicated that the neurocognitive assessment manuals allow for subtest substitutions and indicate which subtests may be used. The assessment manuals indicate that alterations to standard subtest selection and administration should be based on clinical need, not examiner preference.⁴ However, due to lack of protocol-specified standardization and insufficient documentation, it is unclear how examiners selected subtests for administration.

Only one subtest substitution is allowed per the assessment manuals for each index score.⁵

The context of use for neurocognitive assessment data for ongoing subject monitoring in clinical practice and for regulatory decision-making differ in critical ways. In clinical practice, the setting for which neurocognitive assessment manuals are written, flexibility in test administration is acceptable as the clinical care decisions that follow from the test scores likely only impact the one subject and their immediate clinical care decision. Regulatory decision-making is a high-stakes context in which decisions may impact the entire trial for the duration of the trial and can set a medical and scientific precedent for generations.

⁴ E.g., WPPSI-III administration and scoring manual, page 21.

⁵ E.g., WISC-IV administration and scoring manual, page 27.

4. Administration with linguistic interpreters

Neurocognitive assessments were conducted with the assistance of linguistic translators for non-Italian speaking subjects. The presence of a translator is not consistent with standardized administration of these neurocognitive assessments and may impact subject performance, although the impact on a subject's performance is unknown and not necessarily unidirectional. In a neurocognitive report written in Italian provided by the Applicant, the test administrator also noted that the presence of a translator may have impacted the evaluation, but the impacts were not known.

“L'esame viene condotto con l'ausilio di un facilitatore linguistico e accompagnato dall'osservazione clinica e da un colloquio di aggiornamento con il caregiver (madre) allo scopo di indagare possibili variazioni del profilo di funzionamento cognitivo e comportamentale del bambino. Si sottolinea che la somministrazione mediante l'utilizzo della traduzione può aver inficiato, in misura non quantificabile, la valutazione stessa.” Quotation from page 2 out of 2 from the Applicant response to the clinical information request #4 section 5.3.5.1 201222 Neurocognitive Data Clinical Summaries – WISC-IV (0026, received December 22, 2023).

5. No quality review oversight of neurocognitive assessment administration and scoring

Neurocognitive assessments are complex tests that require a high level of specialized training to administer. The Applicant confirmed the three neurocognitive test administrators involved in the study were licensed psychologists in Italy (response to clinical IR #1 received September 28, 2023, STN 0006). Even when administrators are highly skilled, there can be errors in the administration and scoring of these tests⁶ (e.g., due to administrator mistakes, behavioral noncompliance), resulting in non-standardized administration and scores that do not necessarily reflect the cognitive ability of the child. It is recommended that a systematic data review plan is developed and implemented in a clinical trial where neurocognitive assessments are used for efficacy endpoints given that scoring errors can interfere with score interpretability.

In a response to clinical IR #4 (received December 13, 2023, STN 0020), the Applicant provided a worse-case scenario to evaluate the impact of potential administration and scoring errors on the cognitive efficacy endpoint based on severe cognitive impairment defined as a performance standard score of ≥ 55 or death from any cause for treated and untreated subjects.

In a response to clinical IR #4 (received December 15, 2023, STN 0021) the Applicant provided administration forms from the Bayley-III and WPPSI-III. Based on this reviewer's evaluation of the Bayley-3 administrations, 10 of the 32 administrations

⁶ Sanchez et al. Common rater errors for assessment in pediatric rare and orphan disease trials. International Society for CNS Clinical Trials and Methodology (ISCTM) 2018 Autumn Conference poster presentation. https://isctm.org/public_access/Autumn2018/PDFs/Sanchez-Poster.pdf

appeared to follow basal/ceiling rules while 22 did not follow the standardized administration rules (e.g., test discontinuation rules). Of the WPPSI-III administrations, not all administration forms were legible; however, 1 of the 5 administrations appeared to follow basal/ceiling rules while 4 of 5 did not (e.g., test discontinuation rules). A similar trend was observed with other assessment administrations. Failure to follow the discontinuation rules for the neurocognitive assessments likely resulted in lower performance standard scores, which likely biases the Applicant's results away from efficacy and towards a conclusion of severe cognitive impairment.