

An Adapted Clinical Measurement Tool for the Key Symptoms of CLN2 Disease

Journal of Inborn Errors of Metabolism
& Screening
2018, Volume 6: 1–7
© The Author(s) 2018
DOI: 10.1177/2326409818788382
journals.sagepub.com/home/iem



Kathleen W. Wyrwich, PhD¹, Angela Schulz, MD²,
Miriam Nickel, MD² , Peter Slasor, ScD³, Temitayo Ajayi, MD³,
David R. Jacoby, MD, PhD³, and Alfried Kohlschütter, MD²

Abstract

Neuronal ceroid lipofuscinosis type-2 (CLN2) disease is a rare, autosomal recessive, pediatric-onset, neurodegenerative lysosomal storage disease caused by mutations in the *TPP1* gene. Cerliponase alfa (Brineura[®]), a recombinant form of human tripeptidyl peptidase-1, was recently developed as a treatment for CLN2 disease. In clinical trials, the primary end point to evaluate treatment effect was the aggregate score for the motor and language (ML) domains of the CLN2 Clinical Rating Scale, an adaptation of the Hamburg scale's component items that include anchor point definitions to allow consistent ratings in multinational, multisite, clinical efficacy studies. Psychometric analyses demonstrated that the ML score of the CLN2 Clinical Rating Scale and individual item scores are well defined and possess adequate measurement properties (reliability, validity, and responsiveness) to demonstrate a clinical benefit over time. Additionally, analyses comparing the CLN2 Clinical Rating Scale ML ratings to the Hamburg scale's ML ratings demonstrated adequate similarity.

Keywords

CLN2 disease, neuronal ceroid lipofuscinosis type-2, Batten disease, cerliponase alfa, tripeptidyl peptidase-1, enzyme replacement therapy, intracerebroventricular, clinician-reported outcome, Hamburg Motor-Language scale, psychometric

Introduction

The neuronal ceroid lipofuscinoses (NCLs), also known as Batten disease, are discrete genetic lysosomal storage disorders characterized by pathologic autofluorescent cellular deposits, with the clinical hallmarks of progressive seizure disorder, ataxia, motor function loss, dementia, and blindness.^{1–3} Neuronal ceroid lipofuscinosis type-2 (CLN2) disease is a rare, autosomal recessive, predominantly late infantile form of NCL caused by mutations in the *TPP1* gene (also referred to as the *CLN2* gene). The *TPP1* gene encodes tripeptidyl peptidase-1 (TPP1), which cleaves N-terminal tripeptides from polypeptides imported into lysosomes destined for cellular degradation. The TPP1 deficiency manifests as central nervous system neuronal cell dysfunction and death, reflected as a rapidly progressive neurodegeneration and loss of physical function.¹

The clinical course of CLN2 disease is largely predictable.⁴ In most cases, affected children appear healthy until age 2 to 4 years. Initial symptoms are typically new-onset seizures and inability to acquire new language milestones. This is rapidly followed by a period of active loss of function, characterized by ataxia, movement disorders, language regression, and cognitive

impairment. After a period of 2 to 4 years, children become unable to walk and talk and eventually go blind. Severe neurological dysfunction occurs by age 4 to 8 years and death typically occurs by age 8 to 16 years.^{1,2,5,6}

The Hamburg scale was developed to quantify the loss of function that occurs over the clinical course of CLN2 disease.⁴ This 4-item instrument assesses motor function (walking ability), visual function, language, and seizures, with all items using 0 to 3 rating options and yielding a 0 to 12 total summed score. Worgall and colleagues⁵ introduced a second CLN2 disease rating scale, the Weill Cornell Scale, which assesses

¹ Eli Lilly and Company, Indianapolis, IN, USA

² University Hospital Hamburg-Eppendorf, Hamburg, Germany

³ BioMarin Pharmaceutical Inc, San Rafael, CA, USA

Received March 28, 2018, and in revised form June 8, 2018. Accepted for publication June 12, 2018.

Corresponding Author:

Kathleen W. Wyrwich, PhD, Eli Lilly and Company, 3647 Hartford St, St. Louis, MO 63116, USA.

Email: wyrwich_kathy@lilly.com



Table 1. The Hamburg Motor and Language (HML) and CLN2 Clinical Rating Scale Motor-Language (ML) Domain Items.

HML Scale Items in Natural History Registry		ML Scale Items in Cerliponase Alfa Clinical Trials		Rationale for Adaptations
Motor	3 Walks normally ^a	Motor	3 Grossly normal gait. No prominent ataxia, no pathologic falls.	Clarification
	2 Frequent falls, clumsiness obvious	2	Independent gait, as defined by ability to walk without support for 10 steps. Will have obvious instability and may have intermittent falls	Added "step" criteria to clarify and harmonize the definition of gait changes across sites/investigators
	1 No unaided walking or crawling only	1	Requires external assistance to walk or can crawl only	Clarification
	0 Immobile, mostly bedridden	0	Can no longer walk or crawl	No changes
Language	3 Normal (individual maximum) ^b	Language	3 Apparently normal language. Intelligible and grossly age-appropriate. No decline noted yet.	Clarification
	2 Has become recognizably 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).	Added the language baseline definition to loss of function
	1 Hardly understandable	1	Hardly understandable. Few intelligible words	No changes
	0 Unintelligible or no language	0	No intelligible words or vocalizations	No changes

^aIn some children, motor development was never really normal.

^bIn some children, normal language development was never present. In such cases, the best performance ever achieved was taken as a starting point and rated 3; when language then became recognizably worse, it was rated 2.

4 items: gait (walking ability), language, myoclonus (involuntary movements), and feeding/swallowing. Like the Hamburg scale, all Weill Cornell Scale items use 0 to 3 rating options and yield a 0 to 12 total summed score.

Brineura[®] (BioMarin Pharmaceutical Inc., Novato, CA) (cerliponase alfa), a recombinant form of TPP1, has been developed as a treatment for CLN2 disease.⁷ Cerliponase alfa is administered directly into the central nervous system via an indwelling intracerebroventricular port. Determining the clinical efficacy of this new treatment required the use of a clinical outcome assessment (COA) tool capable of demonstrating (1) a clinically relevant shift in the course of disease compared to the natural history for children and adolescents with CLN2 disease and (2) the reliability, validity, responsiveness, and interpretability required for COA use in a multinational clinical trial.⁸ An adapted scale, termed the CLN2 Clinical Rating Scale, was devised to include historically used motor and language (ML) anchor point definitions from existing natural history databases to enable prospective acquisition of clinical trial data comparable to existing historical data.

This report details the key measurement properties of reliability, validity, and responsiveness of the CLN2 Clinical Rating Scale and the 2 component scores (motor and language) for use in CLN2 clinical trials. Moreover, the process and results from testing the similarity with the Hamburg ML (HML) scale are described and reported, thus demonstrating the comparability of the HML scores, derived from a natural history cohort, to the CLN2 Clinical Rating Scale ML scores used in 2 cerliponase alfa clinical trials. This demonstration of comparability is an

important aim of this report and provides assurance that the HML and the CLN2 Clinical Rating Scale ML scores are adequately similar for use in the interpretation of functional change over time in the clinical studies.

Methods

CLN2 Clinical Rating Scale ML Domain

The CLN2 Clinical Rating Scale ML domain was adapted from the common subscales of the Hamburg and Weill Cornell CLN2 clinical rating scales to be used as an assessment tool for multicenter efficacy studies supporting the development of cerliponase alfa. The ML functions are fundamental disease domains, decline rapidly and predictably as a function of age, and are relatively insensitive to standard of care.

The rating is structured so that a score of 3 indicates a *normal condition*, 2 is a *slight or just noticeable abnormality*, 1 is a *severe abnormality*, and 0 denotes a *complete loss of functioning*. Because of wording ambiguity and differences between scales, item wording and anchor point definitions were refined for the CLN2 Clinical Rating Scale. This adaptation was conducted to codify the historical transitions in the items such that the treatment study could be directly compared to natural history gathered at both the University Hospital-Eppendorf in Hamburg and the Weill Cornell Medical College. The changes used in the CLN2 ML scale and the rationale are shown in Table 1.

Differences in 1-point scoring on the ML domains are measurable and clinically important. For example, in the motor

domain, a 1-point drop from a rating of 3 to a rating of 2 is the difference between a child who can walk normally and a child who is abnormal but retains independent ambulation of at least 10 steps. Further, 1-point drop to a score of 1 indicates a child who no longer independently ambulates but can still move by some self-process. The last 1-point drop to a score of 0 indicates a child who is no longer self-mobile. Similarly, it can be observed that each of the language domain ratings also represent clinically meaningful levels.

Patient Population

A dose-escalation, single-arm, multinational phase 1/2 study (Clinicaltrials.gov identifier NCT01907087) evaluated the efficacy and safety of cerliponase alfa in patients with CLN2 disease. It included a stable dose period of 48 weeks at a dose of 300 mg every 2 weeks (Q2W), administered via an intracerebroventricular catheter. An ongoing open-label extension study (Clinicaltrials.gov identifier NCT02485899) continues to evaluate the efficacy and safety of cerliponase alfa. Any patient who completed treatment in the phase 1/2 study was eligible. All patients continue to receive cerliponase alfa at 300 mg Q2W for up to 240 weeks. Patients enrolled in the DEM-CHILD-independent CLN2 natural history registry⁹ served as an untreated comparison group.

Eligible patients were required to have documented *TPP1* deficiency and genotype. Children with CLN2 disease were eligible who were at least 3 years old and disease score ≥ 3 on the 0- to 6-point CLN2 Clinical Rating ML scale, with scores of at least 1 in each of these domains. Written informed consent from a parent/legal guardian and assent from the patient, if appropriate, were obtained. The studies were performed in accordance with the Declaration of Helsinki; approval was obtained from the institutional review board at each participating center.

Interrater Reliability

Clinical ratings were performed in-person every 8 weeks, and each assessment was videotaped. The prespecified methodology was as follows: 36 videotapes from 4 representative and regular time points (baseline, week 25, week 49, and week 73 [week 25 of extension study]) for 11 patients enrolled in the clinical studies who were/are being treated at the Hamburg clinic were independently rated by a nonstudy scale trainer. These ratings were then compared to the motor, language, and combined ML scale ratings given by the respective study clinicians. Weighted kappa (κ_w) statistics¹⁰ assessed the level of rater agreement between the nonstudy trainer's ML scale ratings when compared to the study investigators' ratings to better understand interrater reliability using the Landis and Koch¹¹ classifications of agreement (0.00-0.20 indicates slight agreement, 0.21-0.40 indicates fair agreement, 0.41-0.60 indicates moderate agreement, 0.61-0.80 indicates substantial agreement, 0.81-0.99 indicates almost perfect agreement, and 1.00 indicates complete/perfect agreement).

Construct Validity

The validity of an assessment tool indicates how well it measures what it is designed to measure. Validity is supported when different methods of measuring the same or similar constructs produce similar results. To examine construct validity, the pattern and magnitude of the relationship between the scores for the motor, language, and combined ML scale scores were compared to other similar health measures using Spearman correlation coefficients in the baseline data. For all correlational analyses, the absolute value of the Spearman correlation coefficients assessed whether weak ($|r| < 0.30$), moderate ($0.30 \leq |r| < 0.60$), or strong ($|r| \geq 0.60$) relationships exist.¹²

Similar health measures compared to the motor score included the Physical Functioning Domain (8 items) and the Total Scale score from the 23-item Pediatric Quality of Life (PedsQL) Generic Core Scales, V4.0.¹³ Due to the impact of the child's motor functioning on a parent's caregiving role, the motor scores were compared to 3 key domain scores (Physical Functioning, Emotional Functioning, and Social Functioning) and the Total Score of the PedsQL Family Impact Module.¹⁴ In addition, the Daily Activities Scale score from the CLN2 Quality of Life Questionnaire (CLN2 QL) was also examined to support the construct validity of the motor item's score.

The language score was compared to other PedsQL Family Impact Module Domain scores for Social Functioning, Cognitive Functioning, and Communications reflecting the impact of the child's language abilities on these parenting roles. Finally, the combined ML scale score was compared to the PedsQL Total Score, the PedsQL Family Impact Module Total Score, and the CLN2 QL Daily Activities Score.

Responsiveness

The responsiveness (ie, the ability to detect change over time) of the motor, language, and combined ML scores was also examined using Spearman correlation coefficient. The ML scores for change from baseline to week 49 (Study 201 Completion/Early Termination) were compared to the change score from baseline to the respective visit on the CLN2 QL Daily Activities Score. The absolute value of the Spearman correlation coefficient assessed whether weak ($|r| < .30$), moderate ($.30 \leq |r| < .60$), or strong ($|r| \geq .60$) relationships exist.¹²

Comparison Study of HML and CLN2 Clinical Rating Scale ML Domains

Ensuring the adequate similarity of measurements using the HML and the CLN2 Clinical Rating Scale ML domains is key to the interpretation of clinical trial results. Ideally, all natural history assessments could be scored using the ML domains by clinical trial raters. However, due to the historical nature of these data, subsequent deaths from some of these children, and the fact that videotaping patients were not part of the clinical

acquisition routine in the natural history registry, a sufficient supply of natural history patient videos was not feasible.

Therefore, to bridge the similarities of these 2 rating scales, videotaped assessments were reviewed and scored using the HML scale by a rater independent of the clinical studies, trained on the HML definitions. The goal of this bridging investigation was to understand the degree of variability in ratings in the analyses that is attributable to the slight changes in the study rating definitions. The videos were viewed in their entirety, and the videos were not transcribed for review due to the lack of meaning that comes from the transcription of videos of patients with CLN2, many with quite limited language abilities.

κ_w statistics¹⁰ assessed the level of rater agreement between the HML scale ratings (motor, language, and combined ML scores) when compared to the clinical trial clinician ratings using the ML scale. The κ_w results were interpreted using the Landis and Koch¹¹ classifications of agreement (see “Interrater Reliability” section) to assess the similarity of the 2 rating scales.

In addition, slope estimates for the rate of decline over 48 months of treatment were produced for each of the original and ML scale scores using the same algorithm as specified in the phase 1/2 study. The slope for a patient was computed as a change from baseline, using the last observation >0 and scaled as a rate per units of 48 weeks. The analyses were also produced for the separate ML components. Summary statistics (mean, median, standard deviation [SD]) were produced for the HML and Clinical Rating Scale ML slope estimates.

Results

Patient Population

A total of 24 patients were screened for the phase 1/2 study at 5 clinical sites. All screened patients were enrolled. One patient was discontinued after placement of the intracerebroventricular catheter and the first 300 mg infusion at the parents' request, due to the patient's unwillingness to continue with study procedures. The remaining 23 patients completed the phase 1/2 study and went on to enroll in the extension study.

The mean (SD) age of the study patients at enrollment was 4.3 (1.24) years (Table 2). Most patients were white (96%); 63% were female. The mean (SD) age of CLN2 disease onset was 3.4 (1.07) years (Table 2). One or both of the 2 most common CLN2 alleles (ie, c.622C>T and c.509-1G>C) were present in 17 (71%) patients. The mean (SD) baseline ML scale score was 3.5 (1.20) points. Three patients had baseline ML scale scores below the study enrollment minimum of 3; however, study eligibility was determined at the screening rather than baseline assessment. The protocol permitted a 1-month interval between the screening and baseline visits.

Interrater Reliability

In the formal interrater study of Hamburg clinic patients using patient videos, κ_w results for motor score ratings for all videos

Table 2. Baseline Demographic Characteristics (Phase 1/2 Study, Enrolled Population).

	All Patients, N = 24
Age at enrollment, years	
Mean (SD)	4.3 (1.24)
Median	4.0
Minimum, maximum	3.0, 8.0
Sex, n (%)	
Female	15 (63)
Male	9 (38)
Race, n (%)	
Asian	1 (4)
White	23 (96)
Ethnicity, n (%)	
Hispanic or Latino	1 (4)
Not Hispanic or Latino	23 (96)
Age at disease onset, years, n (%)	
<3	7 (29)
3-<5	12 (50)
≥5	4 (17)
Unknown	1 (4)
Mean (SD)	3.4 (1.07)
Median	3.0
Minimum, maximum	2.5, 6.3
Genotype, n (%)	
c.622C>T	5 (21)
c.509-1G>C	2 (8)
c.622C>T and c.509-1G>C	2 (8)
c.622C>T and other	4 (17)
c.509-1G>C and other	4 (17)
Other	7 (29)
Screening ML Scale score, n (%)	
Mean (SD)	3.7 (0.95)
Median	3
Minimum, maximum	3, 6
Baseline ML Scale score ^a	
Mean (SD)	3.5 (1.20)
Median	3
Minimum, maximum	1, 6

Abbreviations: ML, motor and language; SD, standard deviation.

^aStudy-eligibility criteria were based on the baseline rather than screening visit.

(n = 36) were .93, indicating near-perfect agreement.¹¹ Perfect agreement ($\kappa_w = 1.00$) were observed at week 25 (n = 11), week 49 (n = 10), and week 73 (n = 3). At baseline, 2 ratings were provided that were 1 level lower than the clinical study ratings ($\kappa_w .76$; n = 12), but nonetheless, these ratings demonstrated substantial agreement (Table 3).

Similarly, for the language score ratings across all videos, $\kappa_w = .82$, thus indicating near-perfect agreement. The κ_w estimates at each time point ranged from .67 at week 49 to .93 at baseline. Moreover, the combined ML scale scores (0-6) generated from the ML scale trainer's ratings were in almost perfect agreement with those of the clinician raters ($.89 \leq \kappa_w \leq .93$) for each time point and across all rating periods—supporting interrater reliability for the ML scale scores. Moreover, the demonstrated consistency between the clinical trial ratings made by each of the study clinicians and the study's ML scale

Table 3. Interrater Reliability Estimates^a in Clinical Studies Among Hamburg Clinic Patients.^b

	Baseline, n = 12	Week 25, n = 11	Week 49, n = 10	Week 73, n = 3	All 4 Time Points, n = 36
ML scale motor ratings	0.76	1.00	1.00	1.00	0.93
ML scale language ratings	0.93	0.79	0.67	0.80	0.82
Combined (0-6) ML scale scores	0.92	0.93	0.89	0.93	0.92

Abbreviation: ML, motor and language.

^aWeighted kappa comparison between 2 raters.

^bConducted in 2016.

trainer supports the reproducibility of the ML ratings and procedures (Table 3).

Construct Validity

Construct validity analyses for the motor, language, and combined ML scale scores examined the magnitude of Spearman correlation coefficients at baseline. For the domains and total scores most closely related to the motor score, baseline correlations were generally moderate to strong, with the strongest observed relationships with the PedsQL Core Physical Functioning measure ($r = 0.64$) and Total Score ($r = 0.61$). For the domains scores most closely related to the language score, baseline correlations were generally moderate to strong with the PedsQL Family Impact Module's Communication ($r = .65$), Social ($r = .57$), and Cognitive Functioning ($r = .47$) Domain comparisons. The ML scale score demonstrated moderate relationships with the comparator instruments' Total Scores (PedsQL Total Score, $r = .35$; the PedsQL Family Impact Module Total Score, $r = .51$; the CLN2 QL Daily Activities Score, $r = .35$).

Responsiveness

The ML scale change scores (motor, language, and combined ML) from baseline to week 49 (end of study visit in Study 201) were compared to the change score from baseline to the respective visit on the CLN2 QL Daily Activities Score using the Spearman correlation coefficients. Both the ML and the motor change scores demonstrated a moderate level of responsiveness with the CLN2 QL Daily Activity Score ($r = 0.37$ and 0.41 , respectively). The PedsQL Social Functioning measure's change score correlation with the language change score was 0.40 —also achieving a moderate relationship to demonstrate the responsiveness of the ML combined and standalone items.

Comparison Study of HML and CLN2 Clinical Rating Scale ML Score

A bridging study using videotapes and ratings by an HML rater investigated whether adequate similarities exist between the ML scores of the HML and the CLN2 Clinical Rating Scale. The results demonstrate substantial-to-perfect agreement between the 2 corresponding motor scales ($.67 \leq \kappa_w \leq 1.00$), fair-to-substantial agreement between the language scales ($.34 \leq \kappa_w \leq .62$), and substantial agreement in the

Table 4. Estimated Rate of Decline Over 48 Weeks in Phase 1/2 Study: Motor, Language, and Combined Scores Slope Estimates in the Scale Comparison Study.

Rate of Decline (Points/48 Weeks) ^a	CLN2 Rating Scale	
	ML	HML
Motor scores	n = 12	n = 12
Mean (SD)	0.17 (0.33)	0.14 (0.25)
Median	0.00	0.00
Minimum, maximum	0.00, 1.01	0.00, 0.61
95% CI limit	-0.04, 0.39	-0.02, 0.30
Language scores^b	n = 6	n = 6
Mean (SD)	-0.02 (0.58)	0.00 (0.00)
Median	0.00	0.00
Minimum, maximum	-0.61, 1.01	0.00, 0.00
95% CI limit	-0.62, 0.58	0.00, 0.00
ML combined score	n = 12	n = 12
Mean (SD)	0.17 (0.52)	0.05 (0.41)
Median	0.00	0.00
Minimum, maximum	-0.61, 1.01	-1.00, 0.61
95% CI limit	-0.16, 0.50	-0.21, 0.31

Abbreviations: CI, confidence interval; CLN2, neuronal ceroid lipofuscinosis type-2; HML, Hamburg Motor-Language; ML, motor and language; SD, standard deviation.

^aNegative slopes denote improvement over time.

^bFive patients were omitted from language slope estimation because the baseline language score was 0. One additional patient was omitted because there were no observations postbaseline with a language score >0.

combined ML scores ($.67 \leq \kappa_w \leq .79$) at all of the investigated time points (baseline, week 25, week 49, and week 73). The mean slopes of ratings from baseline to week 49 computed for the HML and the ML scale ratings yielded nearly identical point estimates and variation for the Hamburg clinic patients in this investigation (Table 4). This was true for the mean slopes from the respective motor scales (0.14 ± 0.25 [HML] vs 0.17 ± 0.33 [ML]), language scales (0.00 ± 0.00 [HML] vs -0.02 ± 0.58 [ML]), and 0 to 6 ML combined score from each scale (0.05 ± 0.41 [HML] vs 0.17 ± 0.52 [ML]). These results provide assurance that the HML and the CLN2 Clinical Rating Scale ML scores are adequately similar for use in the interpretation of functional change over time in the clinical studies.

Discussion

Clinical outcome assessments designed to demonstrate the benefits of a treatment must be well defined and possess adequate

measurement properties.⁸ The primary metric used to quantify a notable attenuation of CLN2 disease progression in clinical studies was the aggregate of the ML score in the CLN2 Clinical Rating Scale. In order to ensure measurement consistency across clinical trial sites, the CLN2 Clinical Rating Scale ML domain was adapted from the original Hamburg scale⁴ to include anchor point definitions that would allow consistent ratings within the study conduct. The adapted CLN2 Clinical Rating Scale ML domain was intended to respect the historical application of the Hamburg scale and allow consistent and harmonized ratings in a multinational, multisite, clinical efficacy study.

The psychometric analyses of the ML rating scale with regard to reliability, construct validity, and responsiveness provide confidence in the measurement properties of this adapted instrument and the 2 component scores. Interrater reliability was assessed using Hamburg clinic patients in the clinical studies; the results are convincing and show that agreement for the motor, language, and combined ML scores was substantial to nearly perfect.¹¹ These results reflect the important success in achieving rater consistency in the clinical studies through strong and thorough rater trainings. Construct validity and ability to detect change were also demonstrated with moderate-to-strong correlations with key domains and total scores from quality-of-life instruments (PedsQL, PedsQL Family Impact Module, and CLN2 QL Daily Activities Score) measuring similar constructs.

Agreement with the ML scale, as assessed by κ_w values for single and combined ML scores, was near perfect for the motor score for all time points; there was fair-to-substantial agreement for the language scores.¹¹ External review of videotapes may pose challenges for precise language assessments that may contribute to the lack of strong agreement in the language score. In contrast to study raters, the external reviewer had no previous medical or social information such that previous maximum attained function was not known. Technical issues such as recording quality or sample bias that may limit the ability of the rater to accurately assess language capabilities may also confound results. However, the external video rating assessment of decline using the HML scale led to nearly identical values for progression slope point estimates and variability when compared to the individual and combined domains of the ML scale's rating changes over time, suggesting a strong relationship in the assessment of disease progression despite the lack of strong agreement in language scores. These results provide assurance that the ML scores from the 2 scales are adequately similar for use in the interpretation of the results from the clinical studies when compared to the natural history registry.

This study is not without limitations. First, no other known psychometric validation of this scale has been conducted outside these clinical studies; the use of the adapted instrument is encouraged to better understand the strength and durability of the conclusions in this report in other CLN2 disease investigations. In addition, no formal intrarater reliability study was conducted. However, the correspondence of ratings between

the clinician's rating compared to the ratings by the rater trainer using the videotaped assessments was strong and demonstrated that each study clinician made his or her ratings in a manner that was strongly consistent with the opinion of the trainer on the study rating process—as may be expected by the use of well-trained physicians across these global academic medical centers.

Other limitations included the hampered ability to evaluate responsiveness (the ability to detect change over time) in these analyses given that few ML score shifts were observed in patients in the phase 1/2 study (Table 4). Nonetheless, moderate change correlations ($r \geq 0.37$) with key quality-of-life domains scores provide confidence in this measurement property. Finally, the scales' video-based ratings were subsequently compared to the primary source data for the clinical studies, which is the in-person rating by the patient's Hamburg clinic physician. The use of the videos introduced possibly important sources of error in comparing these ratings; namely, the rater was not present for any interactions between the patient and the study clinician prior to the start of videotaping (eg, patient walks up to greet the study physician in the clinic hallway) or any unobservable (on the video) details (eg, someone walking into the room, pictures in the room or the view outside the window, etc). In contrast, the study clinician often knew each patient was physically present throughout the clinic visit and assessment and was capable of making a comprehensive assessment that may not have been fully captured in the video. This limitation of video assessments may contribute to the lack of strong agreement between the raters' score of language ability, which may be more difficult to capture fully on video than motor ability, which had substantial-to-perfect agreement between the 2 corresponding motor scales. Importantly, this did not limit agreement in the assessment of change in language ability over time.

Conclusions

These results demonstrate that the ML domains of the CLN2 Clinical Rating Scale are objective, reproducible, valid, and responsive measures in children with CLN2 disease. These results also provide assurance that the HML and the CLN2 Clinical Rating Scale ML scores are adequately similar for use in the understanding of results from clinical trials. Moreover, change over time on the ML scores is interpretable.⁸ That is, each unit difference on this scale measures a clinically meaningful change in a child's ML abilities. Indeed, anyone who has personally experienced a child's development knows that a change—especially a decline on this scale—is notable, relevant, and reflective of a meaningful change in a child's functioning in the clinical course of CLN2 disease.

Acknowledgments

The authors thank Janet Wittes at Statistics Collaborative, Inc, for statistical assistance with this research and Drs Lim Ming and Ruth

Williams for their participation in the discussions operationalizing the CLN2 Clinical Rating Scale items.


Declaration of Conflicting Interests

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: Peter Slasor, Temitayo Ajayi, and David R. Jacoby are employees and shareholders of BioMarin Pharmaceutical Inc.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Kathleen Wyrwich was employed of Evidera Inc when this research was conducted, which received financial support from BioMarin Pharmaceutical Inc. in connection with the study development, execution, and manuscript development.

ORCID iD

Miriam Nickel  <http://orcid.org/0000-0002-1870-1142>

References

- Kohlschütter A, Schulz A. CLN2 disease (classic late infantile neuronal ceroid lipofuscinosis). *Pediatr Endocrinol Rev.* 2016; 13(suppl 1):682-688.
- Chang M, Cooper JD, Davidson BL, et al. CLN2. In: Mole SE, Williams RE, Goebel HH, eds. *The Neuronal Ceroid Lipofuscinoses (Batten Disease)*. Oxford: Oxford University Press; 2011: 80-109.
- Mink JW, Augustine EF, Adams HR, Marshall FJ, Kwon JM. Classification and natural history of the neuronal ceroid lipofuscinoses. *J Child Neurol.* 2013;28(9):1101-1105.
- Steinfeld R, Heim P, von Gregory H, et al. Late infantile neuronal ceroid lipofuscinosis: quantitative description of the clinical course in patients with CLN2 mutations. *Am J Med Genet.* 2002;112(4):347-354.
- Worgall S, Kekatpure MV, Heier L, et al. Neurological deterioration in late infantile neuronal ceroid lipofuscinosis. *Neurology.* 2007;69(6):521-535.
- Williams RE, Aberg L, Autti T, Goebel HH, Kohlschütter A, Lonnqvist T. Diagnosis of the neuronal ceroid lipofuscinoses: an update. *Biochim Biophys Acta.* 2006;1762(10):865-872.
- Schulz A, Ajayi T, Specchio N, et al; CLN2 Study Group. Study of intraventricular cerliponase alfa for CLN2 disease. *N Engl J Med.* 2018;378(20):1898-1907.
- Walton MK, Powers JH 3rd, Hobart J, et al; International Society for Pharmacoeconomics and Outcomes Research Task Force for Clinical Outcomes Assessment. Clinical outcome assessments: conceptual foundation-report of the ISPOR clinical outcomes assessment – emerging good practices for outcomes research task force. *Value Health.* 2015;18(6):741-752.
- Schulz A, Simonati A, Laine M, Williams R, Kohlschütter A, Nickel M. Abstracts of the 11th EPNS Congress OP48 – 2893: The DEM-CHILD NCL Patient Database: a tool for the evaluation of therapies in neuronal ceroid lipofuscinoses (NCL). *Eur J Paediatr Neurol.* 2015;19:S16.
- Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol Bull.* 1968; 70(4):213-220.
- Landis JR, Koch GG. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics.* 1977;33(2):363-374.
- Hinkle DE, Jurs SG, Wiersma W. *Applied Statistics for the Behavioral Sciences*. Boston: Houghton Mifflin; 1988.
- Varni JW, Seid M, Kurtin PS. PedsQL 4.0: reliability and validity of the Pediatric Quality of Life Inventory version 4.0 generic core scales in healthy and patient populations. *Med Care.* 2001;39(8): 800-812.
- Varni JW, Sherman SA, Burwinkle TM, Dickinson PE, Dixon P. The PedsQL family impact module: preliminary reliability and validity. *Health Qual Life Outcomes.* 2004;2:55.