Addendum to the Final Statistical Analysis Plan

Protocol Title	An Open-Label Study of the Safety, Tolerability, and Pharmacokinetics of Oral NNZ-2591 in Pitt Hopkins Syndrome
Protocol Number:	NEU-2591-PTHS-001
Protocol Version:	3.0
SAP Version:	1.0
Date:	30APR2024
Sponsor:	Neuren Pharmaceuticals Ltd. Suite 201, 697 Burke Road Camberwell, VIC 3124 Australia
Author:	[redacted] [redacted]

This study will be conducted in compliance with the protocol, Good Clinical Practice and all other applicable regulatory requirements, including the archiving of essential documents.

Signatures / Approvals

DocuSigned by:

[redacted]

Signing Reason: I approve this document

Signing Time: May-21-2024 | 23:11 AEST

084C5CFA7A134B96A098C386EDB1BAB2

May-21-2024 | 23:11 AEST

[redacted] Name (Print or Type) Date

Chief Science Officer

[redacted]

DocuSigned by:

[redacted]

Signing Reason: I approve this document

Signing Time: May-21-2024 | 07:44 EDT

8E8903CC7EAF48B294BFB2DEC7E8BF81

May-21-2024 | 07:44 EDT

[redacted] Name (Print or Type) Date

Chief Medical Officer

[redacted]

DocuSigned by:

[redacted]

Signing Reason: I approve this document

Signing Time: ISDT 22:09 | 20-2024-מאי

5D73CE69A4D74FD695233C009BA7A2DF

May-20-2024 | 22:09 ISDT

[redacted] Name (Print or Type) Date

Vice President, Clinical Development

[redacted]

DocuSigned by:

[redacted]

Signing Reason: I approve this document

Signing Time: May-20-2024 | 14:34 EDT

73017382B73A4B3AA9E4885AEFC3BA77

May-20-2024 | 14:34 EDT

[redacted] Name (Print or Type) Date

Biostatistician

[redacted]

Table of Contents

1 Introduction

This document describes additions and changes to the Final Version 1.0 of the Statistical Analysis Plan (SAP) for Protocol NEU-2591-PTHS-001. The section numbers following the introduction in this document correspond to those updated from the SAP Final Version 1.0; therefore, they are not sequential. The reader is referred to SAP Final Version 1.0 for analysis descriptions for sections that are not included in this document.

5.3 Definitions

- Number of missed doses will be derived based on the number of expected doses – the number of doses taken. If a subject took more doses than expected, the number of missed doses will be taken to be 0.

The reader is referred to Section 5.3 of the SAP Final Version 1.0 for other analysis definitions.

9.2.10 Behavior Problems Inventory-Short Form

The BPI-SF will be completed at Visit 3 (Week 0) and at the End of Treatment (Week 13, Visit 16). The BPI-SF is a caregiver-completed rating scale for assessing self-injury, aggressive and stereotyped behavior. Each item is measured for frequency and items in the Self-Injurious Behavior and Aggressive/Destructive Behavior subscales are also assessed for severity. The frequency is scored on a 5-point scale, where 0=never/no problem; 1=monthly, 2=weekly, 3=daily, 4=hourly. The severity is scored on a 3-point scale, where 1=mild, 2=moderate, 3=severe. The items are grouped in three subscales: Self-Injurious Behavior (score ranges from 0 – 32 for frequency and 8 – 24 for severity), Aggressive/Destructive Behavior (score ranges from 0 – 40 for frequency and 10 – 30 for severity), and Stereotyped Behavior (score ranges from 0 – 48 for frequency). The score for each subscale is the sum of the items within that domain. The total frequency score is the sum of the frequency scores for the three domains. The total severity score is the sum of the severity scores for Aggressive/Destructive Behavior and Self-Injurious Behavior subscales. The total severity and frequency scores will only be calculated if all contributing domain scores are not missing. Missing items will be imputed as defined in Section 5.2.5. A subscale score will only be calculated if all items in that domain are answered or can be imputed. Lower scores indicate lesser frequency and severity in behavior problems. Within subject change is the difference between the score at EOT and the score at baseline.

Summary statistics for the total score, subscale scores and their changes from baseline in the frequency and severity of BPI-SF scores will be displayed at End of Treatment. Within subject changes will be analyzed using a Wilcoxon signed rank test. The change from baseline in the total frequency and severity scores at EOT will be presented in plots as described in Section 9.2. The change from baseline in the subscale scores at EOT will be presented in a forest plot and waterfall plots. The subscale and total scores will be presented in line graphs at each visit.

9.2.12 Gastrointestinal Health Questionnaire Scores

The GIHQ will be completed at all Screening/Baseline period visits (Visits 1, 2, 3), as well as Week 6 (Visit 9), End of Treatment (Week 13, Visit 16), and Follow-up (Visit 17). Each item is scored on its frequency, relevancy, and importance to the caregiver. The Frequency is measured on a 5-point scale where 0=never; 1=almost never; 2=sometimes; 3=often; 4=almost always. The surgery domain records occurrence as 0=No and 1=Yes. The relevancy is scored on a scale from 1 to 4, where 1 is not relevant and 4 is very relevant. The importance is scored on a scale from 1 to 4, where 1 is not important and 4 is very important. Lower frequency scores indicate fewer gastrointestinal problems. For relevancy, higher scores indicate the question is more relevant and for importance, higher scores indicate the question is more important. The total frequency score is the sum of all item frequency scores. Similarly, the total relevancy and total importance scores are the sum of all item relevancy and importance scores, respectively. The items are grouped into 9 subscales. The subscale score is the sum of all items within the subscale, as shown on the CRF. Missing items will be imputed as defined in Section 5.2.5. The subscale scores will only be calculated if all items within the domain are answered or can be imputed. The subscales are as follows:

- General Health/Pain: Items 1 – 5, frequency score ranges from 0 – 20, relevancy score ranges from 5 – 20, importance score ranges from 5 – 20.
- Eating, Chewing, and Swallowing: Items 1 – 9, frequency score ranges from 0 – 36, relevancy score ranges from 9 – 36, importance score ranges from 9 – 36.
- Reflux: Items 1 – 3, frequency score ranges from 0 – 12, relevancy score ranges from 3 – 12, importance score ranges from 3 – 12.
- Gas and Bloating: Items 1 – 5, frequency score ranges from 0 – 20, relevancy score ranges from 5 – 20, importance score ranges from 5 – 20.
- Diarrhea and Constipation: Items 1 – 6, frequency score ranges from 0 – 24, relevancy score ranges from 6 – 24, importance score ranges from 6 – 24.
- Personality and Mood: Items 1 – 5, frequency score ranges from 0 – 20, Relevancy score ranges from 5 – 20, importance score ranges from 5 – 20.
- Medications: Items 1 – 9, frequency score ranges from 0 – 36, relevancy score ranges from 9 – 36, importance score ranges from 9 – 36.
- Surgery: Items 1 – 5, frequency score ranges from 0 – 5, relevancy score ranges from 5 – 20, importance score ranges from 5 – 20.
- Parenting: Items 1 – 8, frequency score ranges from 0 – 24, relevancy score ranges from 8 – 24, importance score ranges from 8 – 24.

Within subject change is the difference between the score at each relevant follow-up visit and the score at baseline. Summary statistics for the change from baseline in the total and each GIHQ domain score will be displayed at all post-baseline visits. Test-retest reliability will be assessed for all scores collected during the Screening period (Visits 1, 2, and 3, pre-dosing) using ICC. Changes from baseline will be analyzed using a Wilcoxon signed rank test at Week 6 and EOT. The change from baseline in the total frequency score at EOT will be presented in plots as described in Section 9.2. The change from baseline in the domain frequency scores at EOT will be presented in a forest plot (with the exception of the Surgery domain) and waterfall plots. The domain and total frequency scores will be presented in line graphs at each visit, with the exception of the Surgery domain, which is scored differently from the other domains and is not comparable.

9.2.16 Bayley Scales of Infant and Toddler Development version 4

The BSID-4 will be performed at Screening (Visit 1) and End of Treatment (Week 13, Visit 16) on those subjects who cannot be assessed using the SB5. It assesses cognitive ability, motor skills, and behavior and provides standardized scores in five domains: Cognitive (COG), Language (LANG), Motor (MOT), Social-Emotional (SOEM), and Adaptive Behavior. The Adaptive Behavior scale was not collected for this study. Domain scores range from 40 – 160, where higher scores indicate greater ability in the domain area of development. Scaled scores, age equivalents and growth scale values (GSVs) can also be calculated for the scale subtests: cognition (CG), receptive communication (RC), expressive communication (EC), fine motor (FM), and gross motor (GM). Scaled scores and age equivalents can be calculated for the social emotional (SE) subtest. Scale subtest scores range from 1 – 19 with a mean of 10 and SD of 3. Age equivalent scores range from < 0:15 – >41:08, Growth scale scores range from 414 – 559 with a mean of 500 and SD of 25. Higher scores indicate greater development in the scale subtest. Raw scores, age equivalent scores and GSV are collected for all subjects. Standard scores and scaled scores are only calculated for participants 42 months and younger.

As part of the analysis, an overall developmental quotient, a verbal developmental quotient, and a non-verbal developmental quotient will also be calculated. These developmental quotients will only be calculated if the contributing scale scores are all non-missing. The developmental quotients will be derived as follows:

Overall Developmental Quotient:

(Mean of the age-equivalent scores in months of the cognition, fine motor, receptive communication, expressive communications scales/chronological age in months) x 100

Verbal Developmental Quotient:

(Mean of the age-equivalent scores in months of the expressive communication and receptive scales/chronological age in months) x 100

Non-verbal Developmental Quotient:

(Mean of the age-equivalent scores in months of the cognition and fine motor scales/chronological age in months) x 100

Subtest/Subdomain Developmental Quotients:

The developmental quotients for the BSID subtests will be derived. The developmental quotients for each subtest are calculated in a similar manner as the overall, non-verbal, and verbal developmental quotients. Subtest Developmental Quotient = (Age Equivalent Score in Months for the subtest / chronological age in months) x 100.

Chronological age in months will be calculated as (date of assessment – date of birth + 1)/30.4375.

For age equivalent values < 2, the result is entered into EDC as 1.9, for analysis. For age equivalent values > X, where X is a numeric value, the result is entered into EDC as X.1, for analysis. For example, ">33" is entered in EDC as 33.1. The age equivalent score entered into EDC will be converted to months for analysis according to the following algorithm:

- (a) Count the numeric portion of the result before the decimal as months.
- (b) Count (the numeric portion of the result after the decimal multiplied by 10) as days and convert to months by dividing by 30.4375
- Age equivalent score (months) = (a) + (b)

The standard scores have a mean of 100 and a standard deviation of 15. The developmental quotient scores are a ratio of developmental age to chronological age and will have a range of 1 to 100. Higher scores indicate a higher IQ or better developmental progress. If any scores are missing, the subject will be excluded from corresponding summaries.

Within subject change is the difference between the score at EOT and the score at baseline. If a score is missing, that subject will be excluded from the corresponding summaries.

Summary statistics for the observed and change from baseline in the age equivalent, growth scale subtest scores, the standard and scale scores (for participants for whom the scores are calculated), and developmental quotients will be displayed at end of treatment. Changes from baseline will be analyzed using a Wilcoxon signed rank test. The change from baseline in the raw scores and GSV at EOT will be presented in plots as described in Section 9.2. The change from baseline in the raw and GSV domain scores will be presented in a forest plot, as described in Section 9.2.

9.2.18 Stanford-Binet Intelligence Scales version 5

The SB5 will be performed at Screening and Visit 16 (End of Treatment). Participants who cannot be assessed on the SB5 due to their developmental level will be assessed using the BSID-4 instead. General cognitive ability is reported as a full scale intellectual quotient score (FSIQ). Cognitive ability in five categories is reported in the index scores: fluid reasoning (FR), knowledge (KN), quantitative reasoning (QR), visual-spatial processing (VS), and working memory (WM). Non-verbal (NVIQ) and verbal intelligence quotients (VIQ) are also recorded. Scores range from 40 to 160, where higher scores indicate a higher IQ. For the analysis of index scores and IQ, the z-deviation method for intellectual disability will be used and z-deviation scores will be derived based on age at assessment according to the published algorithm (Sansone et al. 2014). Standard scores, change-sensitive scores, age equivalents and the raw scores are calculated and entered by the site rater.

For age equivalent values < 2, the result is entered into EDC as 1.9, for analysis. For age equivalent values > x, where X is a numeric value, the result is entered into EDC as X.1, for analysis. For example, ">33" is entered in EDC as 33.1.

Summary statistics on the observed and change from baseline in the z-deviation scores, change-sensitive scores and age equivalents for FSIQ, NVIQ, VIQ. Each category will be displayed at end of treatment. The change from baseline in FSIQ, NVIQ, and VIQ z-deviation scores at EOT will be presented in plots as described in Section 9.2. The change from baseline in the domain z-deviation scores will be presented in a forest plot, as described in Section 9.2.

9.3.2 Correlation of Efficacy Endpoints

The potential covariates in this study are listed in Section 8.2. A 2x2 correlation analysis will be performed on all combinations to evaluate the relationship between the covariates. The Pearson Correlation Coefficient will be used to measure the association between age as a continuous variable and the NVIQ/DQ score as described in Section 9.2. If the coefficient is <.5, we will assume little to no association. The Point-Biserial Correlation Coefficient will also be used to measure the association between combinations of continuous and binary variables such as NVIQ/DQ and ASD diagnosis, respectively. The binary variables will be assigned a numeric value of 1 or 2. The Pearson Chi-Square test will be used to measure the association between combinations of categorical variables such as age group and genotype or sex and history of regression. A p-value < 0.10 would indicate association between the variables.

The correlation analysis may inform the selection of covariates (or combinations thereof) to be used in a linear regression analysis on all pre-specified endpoints. Endpoints of interest for the linear regression analysis will be:

- CGI-I Overall Score
- Caregiver Impression of Change Overall Score
- CGI-S Overall Score
- ORCA t-score
- BPI-SF Total frequency and severity scores
- ABC-2 Total Score
- CSHQ Total Score
- GIHQ Total frequency, relevancy, and importance Scores
- Vineland ABC score
- Vineland GSV, raw, and standard domain scores
- QI-Disability Overall Score
- ICND Score

A multiple linear regression model will be determined using the forward selection method where covariates meeting the entry criterion (p-value <0.2) enter into the model. The final model will include all parameters that have successfully been entered into the model.

[redacted]

[redacted]

[redacted]

[redacted]

Statistical Analysis Plan

Protocol Title	An Open-Label Study of the Safety, Tolerability, and Pharmacokinetics of Oral NNZ-2591 in Pitt Hopkins Syndrome
Protocol Number:	NEU-2591-PTHS-001
Protocol Version:	3.0
SAP Version:	1.0
Date:	30APR2024
Sponsor:	Neuren Pharmaceuticals Ltd. Suite 201, 697 Burke Road Camberwell, VIC 3124 Australia
Author:	[redacted]

This study will be conducted in compliance with the protocol, Good Clinical Practice and all other applicable regulatory requirements, including the archiving of essential documents.

Signatures / Approvals

[redacted] Name (Print or Type) Date

Chief Science Officer

[redacted]

[redacted] Name (Print or Type) Date

Chief Medical Officer

[redacted]

[redacted] Name (Print or Type) Date

Vice President, Clinical Development

[redacted]

[redacted] Name (Print or Type) Date

Biostatistician

[redacted].

Table of Contents

1 List of Abbreviations

ABC-2	Aberrant Behavior Checklist-2
AE	Adverse Event
AMSE	Autism Mental Status Exam
APTT	Activated Partial Thromboplastin Time
ASD	Autism Spectrum Disorder
ATC	Anatomical/Therapeutic/Chemical
AUC	Area under the plasma concentration-time curve
BID	Twice-a-day
BPI-SF	Behavior Problems Inventory - Short Form
CFR	Code of Federal Regulations
CGI-I	Clinical Global Impression - Improvement
CGI-S	Clinical Global Impression - Severity
CRF	Case Report Form
CRO	Contract Research Organization
CSHQ	Child Sleep Habits Questionnaire
CSR	Clinical Study Report
COVID-19	Coronavirus-19
DSM5	Diagnostic and Statistical Manual of Mental Disorders
DSMC	Data Safety Monitoring Committee
ECG	Electrocardiogram
EDC	Electronic Data Capture
EGFR	Estimated Glomerular Filtration Rate
EOT	End of Treatment
ET	Early Termination
FDA	Food and Drug Administration
GCP	Good Clinical Practice
GIHQ	Gastrointestinal Health Questionnaire
GLP	Good Laboratory Practice

GMP	Good Medical Practice
IB	Investigator’s Brochure
ICF	Informed Consent Form
ICH	International Conference on Harmonization
ICND	Impact of Childhood Neurological Disability
IEC	Independent Ethics Committee
IMP	Investigational Medicinal Product
INR	International Normalized Ratio
IRB	Institutional Review Board
ITT	Intent-to-Treat Population
MB-CDI	MacArthur-Bates Communicative Development Inventory
MedDRA	Medical Dictionary of Regulatory Affairs
NVDQ	BSID-4 nonverbal development quotient
NVIQ	SB5 nonverbal z-deviation IQ
PT	Prothrombin Time
QI-Disability	Quality of Life Inventory - Disability
SAE	Serious Adverse Event
SAF	Safety Population
SAP	Statistical Analysis Plan
SB5	Stanford-Binet Intelligence Scales version 5
SD	Standard Deviation
SOC	System Organ Class
SOP	Standard Operating Procedure
ORCA	Observer-Reported Communication Assessment
PK	Pharmacokinetics
PMS	Phelan McDermid Syndrome
PMS-DSRS	PMS Clinician Domain Specific Rating Scale
PP	Per-Protocol Population
TEAE	Treatment-Emergent Adverse Event

TFLs	Tables, Figures, and Listings
TSH	Thyroid-Stimulating Hormone
UR	Urine
WHO	World Health Organization

2 Introduction

Neuren Pharmaceuticals Ltd. is conducting a study to investigate a new treatment for Pitt-Hopkins Syndrome (PTHS). The study background, design and subject assessments for the study are described in the study specific protocol.

The statistical methods to be implemented during the analyses of data collected within the scope of this study will be outlined in this document. The purpose of this plan is to provide specific guidelines from which the statistical analysis will proceed. This statistical analysis plan (SAP) is to be interpreted in conjunction with the protocol. Should the SAP and protocol be inconsistent with respect to the planned analyses, the language of the SAP is governing. Any deviations from this plan will be documented in the clinical study report (CSR).

The statistical principles applied in the design and planned analyses of this study are consistent with the International Conference on Harmonisation (ICH) guidelines E9 (Statistical Principles for Clinical Trials).

2.1 Study Objectives

The primary objective of this study is to investigate the safety, tolerability, and pharmacokinetics of treatment with NNZ-2591 Oral Solution, 50mg/mL Investigational Medicinal Product (IMP) in children and adolescents with Pitt-Hopkins Syndrome.

The secondary objective of this study is to investigate measures of efficacy during treatment with NNZ-2591 Oral Solution, 50mg/mL IMP in children and adolescents with Pitt-Hopkins Syndrome.

2.2 Study Design

This is a Phase 2, open-label study of the safety and tolerability of NNZ-2591 Oral Solution, 50mg/mL IMP in male and female children and adolescents with Pitt-Hopkins syndrome. Approximately 10-20 male and female subjects between the ages of 3 and 17 years were planned to be enrolled and receive treatment of NNZ-2591 Oral Solution, 50mg/mL IMP for a total of 13 weeks.

There will be a total of 17 study visits, comprising 5 in-clinic visits and 12 remote telemedicine/in-home nurse visits. The study will commence with an approximately 4-6 week Screening and Baseline period. During this Screening/Baseline period, subjects are assessed for study eligibility and data is collected to establish the subject’s baseline characteristics and symptom severity using a variety of assessments. Two in-clinic visits and one remote/in-home visit (Visits 1, 2, 3) comprise Screening and Baseline.

Once eligibility is confirmed, subjects are dosed with the starting dose of 4 mg/kg and then be up-titrated to the target dose. Subjects receive IMP for a total of 13 weeks. During the treatment period, there are three in-clinic visits: Week 2 (Visit 5), Week 6 (Visit 9) and Week 13 (Visit 16). There is a combined telemedicine/in-home visit by the site Investigator and an in-home visit by a visiting nurse for safety at Week 1 (Visit 4), Week 4 (Visit 7), Week 8 (Visit 11) and Week 10 (Visit 13). Ideally, these combination telemedicine/in-home visits will be conducted simultaneously whenever possible, or they will be conducted on the same day with the telemedicine visit occurring first. If this is not feasible the in-home and telemedicine components can be conducted over two consecutive days. In-home visits with a visiting nurse will be done at Weeks 3 (Visit 6), 5 (Visit 8), 7 (Visit 10), 9 (Visit 12), 11 (Visit 14), and 12 (Visit 15).

All subjects will also have a combined telemedicine/in-home follow-up visit with the Investigator via telemedicine and an in-person home visit by a visiting nurse approximately 2 weeks after the end of treatment (Week 15, Visit 17).

For all visits designated as in-clinic, an in-person visit is preferred. Off-site assessments may be allowed due to extenuating circumstances due to the COVID-19 health emergency. All requests for off-site assessments must be approved in advance by the Sponsor or Medical Monitor. A combined remote visit or nurse-only in-home visit may be conducted in-clinic by the study staff if agreed upon by the Investigator and the caregiver.

Subjects will be divided into three groups by age for enrollment:

- 13 to 17 years old (Group 1)
- 8 to 12 years old (Group 2)
- 3 to 7 years old (Group 3)

Enrollment commenced with the oldest age group and proceeded as follows. After at least three subjects in Group 1 have received two weeks of treatment at the starting dose, the Data Safety Monitoring Committee (DSMC) will review data on safety and tolerability. If tolerability and safety for those subjects during the specified period is deemed acceptable, enrollment for Group 2 will proceed and dosing will begin at the starting dose. When at least three subjects in Group 2 have received the first two weeks of treatment at the starting dose, the DSMC will review data on safety and tolerability. If tolerability and safety for Group 2 is deemed acceptable, enrollment for Group 3 will commence and dosing will begin at the starting dose.

A summary of the schedule of events and assessments is provided in Table 1 below.

The table of Events and Assessments is from version 2.2 which included ophthalmic exam at week 2 and week 6 (as did version 2.0 and 2.1). This was removed in version 3.0. Since the majority of participants were enrolled on versions 2.0-2.2, and none were enrolled under version 3.0, the 2.2 version has been included in the SAP.

Table 1: Schedule of Events and Assessments for Neu-2591-PTHS

Period	**Screening/ Baseline[a]** **4-6 weeks**		**Baseline**	**Treatment Period**													**Follow-up[b]**
Visit Week	**-4 to -6**	**2 weeks from Visit 1**	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13/EOT[a]**	**15**
Visit Number	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**
Visit window (days)	**±1[a]**	**±1[a]**	**±1[a]**	**±1**	**±2**	**±1**	**±1**	**±1**	**±2**	**±1**	**±1**	**±1**	**±1**	**±1**	**±1**	**-3**	**+4**
Type of Visit	**Clinic**	**Remote and In-Home [c,d]**	**Clinic**	**Remote and In-Home [c,d]**	**Clinic**	**In-Home[d]**	**Remote and In-Home[c,d]**	**In-home[d]**	**Clinic**	**In-Home[d]**	**Remote and In-Home[c,d]**	**In-Home[d]**	**Remote and In-Home[c,d]**	**In-Home[d]**	**In-home[d]**	**Clinic**	**Remote and In-Home [c,d]**
Informed consent	X																
Inclusion/exclusion criteria	X		X														
Medical history	X	X	X														
Confirm documented PTHS diagnosis and genotype	X																

Period	**Screening/ Baseline[a]** **4-6 weeks**		**Baseline**	**Treatment Period**													**Follow-up[b]**
Visit Week	**-4 to -6**	**2 weeks from Visit 1**	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13/EOT[a]**	**15**
Visit Number	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**
Visit window (days)	**±1[a]**	**±1[a]**	**±1[a]**	**±1**	**±2**	**±1**	**±1**	**±1**	**±2**	**±1**	**±1**	**±1**	**±1**	**±1**	**±1**	**-3**	**+4**
PTHS history and exam	X																
Confirm co-morbid psychiatric disorders by DSM-5	X																
Rapid SARS-CoV-2 test	X		X													X	
Autism Mental Status Exam	X															X	
Recent clinical history				X	X		X		X		X		X			X	X
Physical examination[f]	X		X		X		X[g]		X		X[g]		X[g]			X	X[g]

Period	**Screening/ Baseline[a]** **4-6 weeks**		**Baseline**	**Treatment Period**													**Follow-up[b]**
Visit Week	**-4 to -6**	**2 weeks from Visit 1**	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13/EOT[a]**	**15**
Visit Number	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**
Visit window (days)	**±1[a]**	**±1[a]**	**±1[a]**	**±1**	**±2**	**±1**	**±1**	**±1**	**±2**	**±1**	**±1**	**±1**	**±1**	**±1**	**±1**	**-3**	**+4**
Neurological examination	X		X		X		X[g]		X		X[g]		X[g]			X	X[g]
[redacted]	X															X	
Abdominal/ CVA/General Appearance[f]		N		N		N	N	N		N	N	N	N	N	N		N
Ophthalmic examination[h]	X				X				X							X	
Vital signs	X	N	X	N	X	N	N	N	X	N	N	N	N	N	N	X	N
Height/length	X		X													X	
Weight	X	N	X	N	X	N	N	N	X	N	N	N	N	N	N	X	N
Review for AEs/SAEs[i]				X,N	X	N	X,N	N	X	N	X,N	N	X,N	N	N	X	X,N

Period	**Screening/ Baseline[a]** **4-6 weeks**		**Baseline**	**Treatment Period**													**Follow-up[b]**
Visit Week	**-4 to -6**	**2 weeks from Visit 1**	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13/EOT[a]**	**15**
Visit Number	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**
Visit window (days)	**±1[a]**	**±1[a]**	**±1[a]**	**±1**	**±2**	**±1**	**±1**	**±1**	**±2**	**±1**	**±1**	**±1**	**±1**	**±1**	**±1**	**-3**	**+4**
Caregiver Diary (dosing, conmeds, voids, seizures, changes in symptoms/ tolerability)	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Review Dosing Compliance				X,N	X	N	X,N	N	X	N	X,N	N	X,N	N	N	X	X,N
In-clinic administration of first dose[j]			X														
12-lead electrocardiogram (ECG)	X		X		X				X							X	N
CBC with differential	X		X		X		N		X		N		N			X	N

Period	**Screening/ Baseline**[a] **4-6 weeks**		**Baseline**	**Treatment Period**													**Follow-up**[b]
Visit Week	**-4 to -6**	**2 weeks from Visit 1**	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13/EOT**[a]	**15**
Visit Number	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**
Visit window (days)	**±1**[a]	**±1**[a]	**±1**[a]	**±1**	**±2**	**±1**	**±1**	**±1**	**±2**	**±1**	**±1**	**±1**	**±1**	**±1**	**±1**	**-3**	**+4**
Coagulation	X		X		X		N		X		N		N			X	N
Comprehensive Metabolic Panel[k]	X		X	N	X	N	N	N	X	N	N	N	N	N	N	X	N
Urinalysis[l]	X		X	N	X	N	N	N	X	N	N	N	N	N	N	X	N
Urinalysis-Drugs of Abuse[l]	X																
TSH, Free T3, Free T4	X															X	
HbA1c	X															X	
Serum pregnancy test[m]	X								X							X	
Blood samples for					X				X							X	

Period	**Screening/ Baseline[a]** **4-6 weeks**		**Baseline**	**Treatment Period**													**Follow-up[b]**
Visit Week	**-4 to -6**	**2 weeks from Visit 1**	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13/EOT[a]**	**15**
Visit Number	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**
Visit window (days)	**±1[a]**	**±1[a]**	**±1[a]**	**±1**	**±2**	**±1**	**±1**	**±1**	**±2**	**±1**	**±1**	**±1**	**±1**	**±1**	**±1**	**-3**	**+4**
pharmacokinetics																	
Blood sample for biomarkers	X		X													X	
Stool sample for microbiome	X		X													X	
CGI-I									X							X	X
CGI-S	X		X						X							X	X
Stanford-Binet Intelligence Scales version 5 (SB5) or Bayley Scales of Infant Development ver 4[n]	X															X	

Period	**Screening/ Baseline[a]** **4-6 weeks**		**Baseline**	**Treatment Period**													**Follow-up[b]**
Visit Week	**-4 to -6**	**2 weeks from Visit 1**	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13/EOT[a]**	**15**
Visit Number	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**
Visit window (days)	**±1[a]**	**±1[a]**	**±1[a]**	**±1**	**±2**	**±1**	**±1**	**±1**	**±2**	**±1**	**±1**	**±1**	**±1**	**±1**	**±1**	**-3**	**+4**
MB-CDI			X													X	
ORCA			X													X	
Caregiver Top 3 Concerns	X	X	X						X							X	X
ABC-2	X	X	X						X							X	X
Behavior Problems Inventory-Short Form (BPI-SF)			X													X	
Caregiver Impression of Change																X	
Quality of Life Inventory-Disability (QI-Disability)			X													X	

Period	**Screening/ Baseline**[a] **4-6 weeks**		**Baseline**	**Treatment Period**													**Follow-up**[b]
Visit Week	**-4 to -6**	**2 weeks from Visit 1**	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13/EOT**[a]	**15**
Visit Number	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**
Visit window (days)	**±1**[a]	**±1**[a]	**±1**[a]	**±1**	**±2**	**±1**	**±1**	**±1**	**±2**	**±1**	**±1**	**±1**	**±1**	**±1**	**±1**	**-3**	**+4**
Impact of Childhood Neurological Disability-Overall quality of life rating			X													X	
CSHQ	X	X	X						X							X	X
Vineland Adaptive Behavior Scales-3			X													X	
GIHQ	X	X	X						X							X	X
Modified two-minute walk test	X		X													X	
Study Exit Form																	X[o]

X = Conducted by site personnel or caregiver as applicable
N = Conducted by visiting nurse
Abbreviations: ABC-2, Aberrant Behavior Checklist-2; AE, adverse event; Behavior Problems Inventory-Short Form; BPI-SF; CBC: Complete Blood Count; CGI-I, Clinical Global Impression Scale-Improvement; CGI-S, Clinical Global Impression Scale-Severity; CMP: Comprehensive Metabolic Panel; CSHQ, Child Sleep Habits Questionnaire; DSM-5, Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition; EOT, end of treatment; GIHQ, GI Health Questionnaire; ICND, Impact of Childhood Neurological Disability; IMP, Investigational Medicinal Product; MB-CDI, MacArthur-Bates Communicative Development Inventory; ORCA, Observer-Reported Communication Assessment; PTHS, Pitt-Hopkins syndrome; QI-Disability, Quality of Life Inventory- Disability; SAE, serious adverse events; Stanford Binet Intelligence Scales ver 5

[a] The Screening/Baseline period will be conducted over 4-6 weeks. The second visit during the Screening period (Visit 2) should be conducted 2 weeks after the Screening visit (Visit 1) ± 1 day. The subject will then remain in Screening for an additional 2-4 weeks. The caregiver diary must be completed daily during the entire Screening period whether it is 4 weeks or 6 weeks. The Screening visit (Visit 1) and Baseline visit (Visit 3) may be conducted over two days. It is preferrable that ECGs, safety labs and biomarker sample collection are conducted on the same day. No assessments can be conducted until after informed consent has been given by the caregiver or legal guardian. Also, training for the caregivers on all caregiver completed assessments must be done before the caregiver rater may complete the assessments. All assessments must be done within the two-day window for the visit (visit day + or – 1 day). For Visit 3, all assessments must be done before administration of the IMP.

[b] Also for early termination. Subjects who have received IMP for longer than 2 weeks and discontinue prematurely any time after Visit 5 should return to the investigational site for final safety and efficacy assessments as scheduled for the End of Treatment visit (Visit 16) and complete the study exit form.

[c] This is a combined remote and in-person nurse visit. The remote visit is conducted by the site via telemedicine or video facilities as appropriate. Any caregiver-completed assessments should be completed the same day of the visit or no earlier than the day before the visit (except the Caregiver Diary which is completed daily) and reviewed by the Investigator. A modified version of the physical and neurological exam will be done at the remote visits based on the guidance document provided to the Investigator. At Week 4, the site will inform the caregiver of the dose titration. See Section 9.2.

[d] In-home visits will be conducted by a visiting nurse. At every nurse visit, the nurse will collect urine, vital signs including weight, and blood samples for a comprehensive metabolic panel (CMP). They will also conduct an exam assessing general appearance, abdominal/costovertebral angle tenderness, and edema. A suprapubic exam for bladder palpation may be done if decreased urine output is reported. The visiting nurse will also review the dosing diary, query oral intake, and do a review of tolerability. Blood samples for CBC with differential and coagulation will be collected at Weeks 4, 8, 10 and Week 15. At the post-treatment follow-up visit at week 15, the visiting nurse will also collect ECG, in addition to the above assessments. At Week 7, the site will also have a phone call with the caregiver to inform him/her of the dose titration. The site will also inform the visiting nurse so the dosing schedule can be reviewed with the family during the in-home visit. See Section 9.1 and Section 9.2.

[e] A clinic visit is preferred but an off-site visit may be conducted for a scheduled in-clinic visit in extenuating circumstances due to the COVID-19 health crisis. All requests for off-site visit assessments due to extenuating circumstances must be approved in advance by the Sponsor or Medical Monitor.

[f] The in-clinic physical exam will include an abdominal exam and an assessment of costovertebral angle tenderness (CVA). A suprapubic exam for bladder palpation may be done if decreased urine output is reported. General awareness will be captured as part of the neurological exam. For in-home nursing visits, the nurse will do an exam assessing general appearance (e.g., alertness), an abdominal/CVA, and edema. A suprapubic exam for bladder palpation may be done if decreased urine output is reported.

[g] A modified version of the neurological and physical exam will be done for the remote visit based on the guidance document provided to the Investigator.

[h] A complete ophthalmological exam will be conducted by an ophthalmologist at Screening (Visit 1) and End of Treatment (Week 13, Visit 16). Abrief exam (undilated fundoscopy exam) will be done by the ophthalmologist at Week 2 and Week 6. See section 9.5.7).

[i] The visiting nurse will query any changes in tolerability during the in-home visit. Any changes will be noted and provided to the Investigator. The Investigator will also query changes in tolerability and note any changes in the remote assessment. The Investigator is responsible for determining and reporting any changes that qualify as Adverse Events.

[j] The first dose of IMP will be administered in the clinic after all Baseline assessments are completed, or, if the Investigator judges that it is too late in the day, on the following day. The subject should be monitored for at least three hours post first dose. Day 1 of dosing is defined by the date when the first dose is taken. At the visit, the site will dispense an adequate supply of IMP for dispensing by the caregiver at home.

[k] A Comprehensive Metabolic Panel will be assessed every post-enrollment visit (see Section 9.5.2). If abnormal findings are observed on renal or liver function tests, additional tests as is clinically indicated may be conducted as described in Sections 9.5.2.1 and 9.5.2.3).

[l] Drugs of abuse will be evaluated at Screening (See Section 9.5.2).

[m] For female subjects who have reached menarche.

[n] The Bayley is done for subjects who cannot obtain a basal score on SB5.

[o] The study exit form should also be completed if the subject terminates the study early.

2.2.1 Study Treatment

The total administration period for NNZ-2591 Oral Solution, 50 mg/mL Investigational Medicinal Product (IMP) is 13 weeks. Following up-titration, the IMP is administered open-label at the target dose of 12 mg/kg BID.

The DSMC reviews the clinical safety data prior to any individual dose titration and prior to initiation of subsequent age groups. For dose titration, after a subject has had two weeks of dosing at the dose level, the data is reviewed by the DSMC during which time the subject stays on their current dose. As outlined below in Table 2, the IMP is administered at a starting dose of 4 mg/kg BID for up to 3 weeks. Following DSMC review and approval of each dose titration, subjects are up-titrated to 8 mg/kg BID for up to 3 weeks, and then to the 12 mg/kg BID dose. The subject then continues on the assigned dose (or highest dose tolerated) for the remainder of the 13 weeks.

Table 2: Dose Titration Schedule

Week 1 and Week 2	**Week 3**	**Week 4 and Week 5**	**Week 6**	**Week 7 and Week 8**	**Week 9**	**Week 10 to Week 13**
4 mg/kg BID (8 mg/kg daily)	4 mg/kg BID (8mg/kg daily)[1]	8 mg/kg BID (16 mg/kg daily)	8 mg/kg BID (16 mg/kg daily)[1]	12 mg/kg BID (24 mg/kg daily)	12 mg/kg BID (24 mg/kg daily)	12 mg/kg BID (24 mg/kg daily)
	DSMC Review Data Prior to Up-Titration[2]		**DSMC Review Data Prior to Up-Titration[2]**		**DSMC Reviews Data After 2 weeks on 12 mg/kg to Confirm Safety[2]**	

[1] Subjects will stay on their current dose until the data is reviewed by the Data Safety Monitoring Committee (DSMC)

[2] Expected time for DSMC review will be one week.

Subjects who are unable to tolerate the study treatment may be discontinued from the study.

3 Analysis Populations

Intent to Treat Population (ITT): The Intent to Treat population will consist of all subjects deemed eligible and enrolled into the study.

Safety Population (SAF): The Safety population will consist of all subjects in the ITT population who received at least one dose of IMP.

Modified Intent to Treat Population (mITT): The modified Intent to Treat population will consist of all subjects in the ITT population who completed treatment and completed the EOT visit.

The pharmacokinetic analyses will be described in a separate analysis plan and summarized separately.

4 Data Monitoring

An independent DSMC will monitor the progress of the trial and ensure that the safety of trial subjects is not compromised. The DSMC will review the clinical safety data prior to any individual dose titration and prior to initiation of subsequent age groups. The DSMC will review safety data and make an adjudication on: 1) the ability for any individual subject to escalate to the next higher dose and 2) the ability for enrollment to begin in the younger age groups (Groups 2 and 3). The DSMC consists of a clinical chair (pediatric cardiologist), and two physicians experienced in clinical trials with pediatric expertise (but not participating in this study). The DSMC Charter details the DSMC processes, duties, and responsibilities as well as the study-wide stopping criteria.

Data are provided to the DSMC as raw data listings for individual titration and age group reviews, as specified in the DSMC Charter. Additional summary analyses may be requested. For quarterly data reviews safety listings are provided to the DSMC. There is no separate Statistical Analysis Plan for the DSMC; the DSMC received as subset of the safety TFLs as required for the quarterly data reviews.

5 Statistical Methods

The statistical analyses will be reported using summary tables, listings, and figures (TLFs). Numbering for TLFs will be based on the recommended numbering convention provided by the International Conference on Harmonization (ICH).

The analyses described in this plan are considered *a priori*, in that they have been defined prior to database lock. Any analysis added after the database lock will be considered exploratory. Post hoc analyses will be labeled as such on the output and identified in the CSR.

Unless noted otherwise, all statistical tests will be two-sided with a significance level of α = 0.05 and descriptive in nature. Continuous variables will be summarized with means, standard deviations, medians, minimums, and maximums. Categorical variables will be summarized by counts and percent of subjects in corresponding categories. Missing values are not considered for percent calculations, unless stated otherwise. Footnotes may be used to specify the percent calculation as appropriate. Baseline for all analyses is defined in Section 5.2.3.

Individual subject data obtained via the electronic data capture (EDC) system and from external vendors (excluding PK data) will be presented in listings.

All analyses and tabulations will be performed using SAS version 9.4 or higher. Tables, listings, and figures will be presented in portable document format (PDF) and in Rich Text Format (RTF). All analyses will be subject to formal verification procedures. Specifically, results will be verified by independent programming. All documents will be reviewed by the lead statistician to ensure accuracy and consistency of analyses. The validation process will be used to confirm that statistically valid methods have been implemented and that all data manipulations and calculations are accurate. Checks will be made to ensure accuracy, consistency with this plan, consistency within tables and consistency between tables and corresponding data listings.

5.1 Determination of Sample Size

Sample sizes were estimated for this open-label study using within-subject change as the measure of improvement. Given the limited data from interventional trials in these patient populations and that this is the first clinical study of NNZ-2591 IMP in this population, treatment change estimates were based on Clinical Global Impression Scale of Improvement (CGI-I) data from completed clinical trials in other rare, neurodevelopmental disorders. Based on the information provided in Glaze et al. 2019 (Double-blind, randomized, placebo-controlled study of trofinetide in pediatric Rett syndrome), the standard deviation of the CGI-I is approximately 0.7 at a single timepoint. For a sample size of 10 subjects, a two-sided significance level of 0.05, and a Wilcoxon signed-rank test, the study will be able to detect a difference of 0.87 units of change in the CGI-I at a power of 0.90. For a power of 0.80, 0.85, and 0.90, the detectable difference would be 0.75, 0.80, and 0.87 or larger, respectively. The effect size for a power of 0.80, 0.85, and 0.90 is 1.07, 1.14, and 1.24, respectively. If the observed standard deviation is larger, the detectable difference would also be larger. For example, if the standard deviation is 1.5, the detectable difference for a power of 0.80, 0.85, and

0.90 is 1.61, 1.71, and 1.86 or larger, respectively. For a sample size of 20 subjects, a two-sided significance level of 0.05, and a Wilcoxon signed-rank test, the study will be able to detect a difference of 0.55 units of change in the CGI-I at a power of 0.90. For a power of 0.80, 0.85, and 0.90, the detectable difference would be 0.48, 0.51, and 0.55 or larger, respectively.

5.2 Handling of Data

5.2.1 Imputation of Missing/Incomplete Dates

An incomplete date is any date for which either the day, month or year is unknown, but all three fields are not unknown. An incomplete date occurs when the exact date an event occurred or ended cannot be obtained from a patient. For many of the analyses, a complete date is necessary to determine if the event should be included in the analysis (i.e., if the event is treatment-emergent or a medication is concomitant). In such cases, incomplete or missing dates will be imputed.

For incomplete or missing onset dates related to adverse events and concomitant medications, the dates will be imputed as follows:

- For an incomplete start date with missing day but month and year are present: If the reported month and year are the same as Day 1 month and year, then the date will be imputed as the date of Day 1. Otherwise, the start day will be imputed to the first day of the month.
- For an incomplete start date with missing day and month but year is present: If the reported year is the same as Day 1 year, then the date will be imputed as the date of Day 1. Otherwise, the start day will be imputed to January 1.
- If the reported date is completely missing, it will be imputed as the date of Day 1.

For concomitant medications, end dates will be imputed as follows:

- For an incomplete end date with missing day but month and year are present: The reported date will be imputed as the last day of the month.
- For an incomplete end date with missing day and month but year is present: If the reported year is the same as the study completion year, then the date will be imputed as the date of study completion. Otherwise, the end day will be imputed to December 31.
- If the date is completely missing, the event will be imputed as the date of study completion.

If any imputed date causes the end date to occur prior to the start date of the event, the start date of the event will be used for the imputation of the end date. If the imputed date is later than the date of study withdrawal, then the date of study withdrawal will be imputed for the date.

If the date of last dose is missing, then it will be imputed to the subject participation end date.

Dates related to efficacy measures will not be imputed.

In subject data listings, start and stop date of events will be displayed as reported on the eCRF (i.e., imputed values will not be listed).

5.2.2 Imputation of Alphanumeric Data

Should there be instances where a clinical laboratory parameter is reported with embedded non-numeric characters, for example, "<0.1" or ">10", the data will be imputed for quantitative summaries. The actual values as reported in the database will be presented in data listings.

For incorporation in quantitative summaries, the following imputation rule will be employed:

- For values reported as less than a lower limit of detection, the value will be imputed as half of the numeric result preceded by the '<' symbol. For example, "<0.1" will be imputed to 0.05.
- For values reported as greater than an upper limit of detection, the value will be imputed as the numeric result preceded by the '>' symbol. For example, ">10" will be imputed as 10.

5.2.3 Visit Windows

Unscheduled assessments that occur on the same day as a scheduled visit will be considered with the corresponding scheduled visit for analysis. Otherwise, unscheduled assessments will be considered with the most recently completed scheduled visit for analysis. If multiple assessments are windowed to the same visit for analysis, the results from the latest assessment will be used, as repeats may have been done to confirm abnormalities, per the protocol. Any unscheduled visits occurring during the Screening period will not be mapped to a scheduled visit but will be considered in the derivation of the Baseline value.

5.2.4 Baseline Values

Baseline values that are missing will not be imputed for safety or efficacy measures. For efficacy assessments scheduled to be collected at more than one visit during the Screening/Baseline period, the baseline value will be based on the average of all non-missing values from assessments conducted prior to the first dose of study drug. For assessments scheduled to be collected once during the Screening/Baseline period, the value from the most recent non-missing assessment prior to the first dose will be used. For safety and other assessments, the value from the most recent non-missing assessment prior to the first dose will be used.

5.2.5 Post-Baseline Values

For subjects who discontinue the study, data will be analyzed up to the last observation for safety and efficacy measures. Missing data for safety assessments will not be imputed. For efficacy rating scale measures other than the exceptions listed below, if an assessment has been completed but has missing items, missing values shall be imputed as follows:

- If 3 or fewer items are missing, the last observation for that particular data element (e.g., item response) will be carried forward.
- The scoring algorithms will be applied including the observed values and last observation carried forward (LOCF) for the missing items

These rules will only apply to individual items. Total scores, subdomain and subscale scores will not be carried forward. If the assessment is completely missing, items will not be imputed.

These imputation rules do not apply to the CGI-I, Clinical Global Impression Scale – Severity (CGI-S) or Caregiver Impression of Change assessments. If a particular domain is missing for these assessments, the participant will be excluded from the corresponding analysis. These imputation rules also do not apply to the Caregiver Top 3 Concerns assessment. If a concern is missing for this assessment, the average severity will still be calculated from all non-missing concern severity scores.

5.3 Definitions

- Age will be derived as the integer value of (informed consent/assent date - date of birth + 1)/365.25.

- The age at time of assessment will be derived as the integer value of (date of assessment – date of birth + 1)/365.25.

- Body Mass Index (BMI) will be derived as weight in kilograms divided by height in meters squared.

- The change from Baseline measurement for a variable at a particular visit will be derived as the measurement taken at the specified visit minus the baseline measurement. For the Clinical Global Impression Scale – Improvement and Caregiver Impression of Change, change from Baseline is the value at the visit.

- Treatment duration in days will be derived as follows: Date of Last Dose – Date of First Dose + 1.

- The number of expected doses will be derived as (2*treatment duration in days)-1, as subjects are expected to receive only their morning dose on the last day of treatment

- The number of doses taken will be derived as the number of doses taken as indicated by the Administration of First Dose CRF and by the Daily Dosing Log CRF for remaining doses, based on completed date and dose fields

- Number of missed doses will be derived based on the number of expected doses – the number of doses taken

- Study Day will be derived as Date of Event – Date of First Dose (+1 if the date of event ≥ date of first dose)

- Screening/baseline duration in days will be derived as follows: Date of First Dose – Date of Screening

- Study duration will be defined as the total number of days in the study (date of last visit – date of informed consent + 1).

- History of developmental regression will be determined from the PTHS diagnosis page of the CRF question: “Has the child undergone a regression of skills in the past?” If coded as “yes” the subject is considered as having had a past developmental regression.

- Autism Spectrum Disorder (ASD) diagnosis will be derived from the PTHS diagnosis page of the CRF. If ASD is checked under “confirmation of psychiatric co-morbidities”, the subject is considered as having the disorder.

- Bipolar Disorder will be derived from the PTHS diagnosis page of the CRF. If “Bipolar disorders” is checked under “confirmation of psychiatric co-morbidities”, the subject is considered as having the disorder.

- Attention Deficit Disorder will be derived from the PTHS diagnosis page of the CRF. If "Attention Deficit Disorder" is checked under "confirmation of psychiatric co-morbidities", the subject is considered as having the disorder.

- Obsessive-compulsive Disorder will be derived from PTHS diagnosis page of the CRF. If "Obsessive-compulsive disorder" is checked under "confirmation of psychiatric co-morbidities", the subject is considered as having the disorder.

- There will be three age group definitions use for the analysis and applied to all analyses requiring an analysis by age group:
  - Enrollment Age Group
    - 13 to 17 years old
    - 8 to 12 years old
    - 3 to 7 years old
  - Child/Adolescent Age Group
    - 13 to 17 years old
    - 3 to 12 years old
  - Divided Age Group
    - 16 to 17 years old
    - 13 to 15 years old
    - 8 to 12 years old
    - 3 to 7 years old

- Dose modification is defined as dose increased, dose reduced, or dose interrupted responses for the action taken with study treatment on the AE CRF.

- TEAEs are defined as those adverse events that occurred after the administration of first dose of study drug.

- [redacted]

- Improvement in the Top 3 concerns will be derived as a decrease in severity (change from baseline < 0). This is calculated for the average score and for individual concerns by the symptom domains as defined in Section 9.2.8.

- The baseline seizure frequency will be determined based on the number of seizures recorded in the diary during the 4–6-week screening period. The baseline monthly seizure frequency will be derived by dividing the total number of seizures during the screening/baseline period by the screening/baseline duration in months (screening/baseline duration divided by 30.4375). During the treatment period (Visits 4 – 6), the seizure frequency will be derived by summing the number of seizures recorded in the diary over treatment. The monthly seizure frequency will be derived by dividing the total number of seizures during the treatment period by the treatment duration in months (treatment duration divided by 30.3475).

- The baseline major breathing episode frequency will be determined based on the number of major breathing episodes recorded in the Major Respiratory Events log of the CRF during the 4-6 week screening period. The baseline monthly major breathing episode frequency will be derived by dividing the total number of major breathing episodes during the screening/baseline period by the screening/baseline duration in months (Screening/baseline duration divided by 30.4375). During the treatment period (Visits 4 – 16), the major breathing episode frequency will be derived by summing the number of major breathing episodes recorded in the diary over treatment. The monthly major breathing episode frequency will be derived by dividing the total number of major breathing episodes during the treatment period by the treatment duration in months (treatment duration divided by 30.3475).

- Prior medications are those medications started prior to and stopped before the first dose of study drug. Concomitant medications include: 1) medications started prior to, but still ongoing as of the first dose of study drug, 2) medications with a start date on or after first dose date.

- The number of concomitant medications will be summed for each subject for all unique medications that are taken while on treatment. Medications taken with a frequency of PRN or “as needed” will be excluded from this count.

6 Subject Information

6.1 Subject Disposition

Information regarding subject disposition will be summarized for all subjects by age groups as defined in Section 5.3 and overall.

Summaries will include: number of subjects enrolled, number of subjects who screen failed, number of subjects in the ITT, Safety, and mITT populations, number of subjects who complete the study, and number of subjects who discontinue the study early. For those who discontinue early, the primary reason for discontinuation will be summarized. The primary reason for discontinuation from the CRF will be mapped to the Completion/Reason for Non-Completion SDTM Terminology for presentation in the table. A data listing by subject will also be generated.

6.2 Protocol Deviations

Protocol deviations will be collected throughout the duration of the study. Protocol deviations will be assigned a sponsor-defined category type. Each deviation will be defined as important or not important. In addition, a by-subject listing of all protocol deviations (important and not important) will be produced.

6.3 Demographic and Baseline Characteristics

Demographics variables will include: age, age groups, sex, ethnicity, race, height at baseline, weight at baseline, and BMI at baseline. Age and BMI will be derived as specified in Section 5.3.

Baseline characteristics will include: PTHS Variant Types (as recorded on the CRF), history of developmental regression, Autism Mental Status Exam, seizure types at screening (as recorded on the CRF), presence of co-morbid psychiatric disorders (e.g. Autism Spectrum Disorder (ASD), bipolar disorders, attention deficit disorder, obsessive-compulsive disorder), and Stanford-Binet Intelligence Scales version 5 (SB5) NVIQ standard and z-deviation score or the Bayley Scales of Infant Development – Version 4 (BSID-4) NVDQ for subjects who cannot basal on the SB5 or for whom the SB5 is not developmentally appropriate. If the SB5 was attempted but the child was unable to basal on the assessment, the BSID-4 should be administered. If partial scores for the SB5 are entered in the EDC in addition to the BSID-4, only the BSID-4 scores will be summarized. If the child has complete sets of scores on the SB5 and the BSID-4, the scores for the assessment that was captured at both Screening and EOT will be summarized. If a participant has a complete set of scores for the SB5 and BSID-4 at both Screening and EOT, these cases will be adjudicated by the Sponsor before database lock.

Demographics and baseline characteristics will be summarized for all analysis populations (ITT Population, Safety Population, mITT Population).

Subjects will be evaluated for PTHS at the Screening Visit. The diagnosis and genotype will be documented. PTHS genotypes will be summarized and listed by subject.

History of developmental regression will be determined from the question as reported on the PTHS diagnosis page in the CRF: “Has the child undergone a regression of skills in the past?" If coded as "yes" the subject is considered as having had a past developmental regression.

The Autism Mental Status Exam (AMSE) will be performed at the Screening Visit. The AMSE is an 8-item observational assessment used to document communicative and behavioral functioning in subjects with ASD. Each question is scored as 0, 1, or 2. The range of the AMSE score is 0 – 14. Higher scores indicate greater severity of autistic symptoms. A question assessing overall impairment is also completed evaluated on a 4-point scale: No impairment, Mild, Moderate and Severe.

The scoring algorithm is below:

- Question 1-4, Eye Contact, Interest in Others, Pointing Skills, Language: each response is assigned a score of 0, 1, or 2 in the order they appear on the CRF.
- Question 5 Pragmatics of Language: Not Impaired or Not Applicable=0, Reported=1, Observed=2
- Question 6 Repetitive Behaviors/Stereotypy: None=0, Insists on routines=1, any other response=2
- Question 7-8 Unusual or Encompassing Preoccupations, Unusual Sensitivities: None=0, Reported=1, Observed=2

The AMSE score is the sum of the individual question scores above and will only be calculated if no items are missing. The AMSE score and the overall impairment will be summarized.

Seizure type ("Clinical Type" in Seizures and Spells Log form on the CRF recorded at Baseline) and characteristics of seizures ("Description in Seizure Types at Screening" form on the CRF: change in awareness, loss of urine or bowel control, loss of ability to communicate, automatic repeated movements, rhythmic jerking muscle, falls if standing or sitting, head drops, warning before seizure occurred, common triggers, muscle stiffness, muscle twitch, other) are recorded at the Screening Visit for subjects. These descriptive characteristics will be summarized using frequencies and percentages.

The presence of other co-morbid psychiatric disorders will be recorded on the PTHS Diagnosis page of the CRF. If the disorder is checked on the CRF, it will be considered as present in the subject.

6.4 Medical History

Medical history will be coded using Medical Dictionary for Regulatory Activities (MedDRA) version 24.0 and listed by System Organ Class (SOC), Preferred Term (PT), and verbatim term. A complete medical history will be recorded at all Screening and

Baseline visits (Visits 1, 2, 3). Medical history will be summarized using frequencies and percentages according to the SOC and PT.

7 Safety Analysis

The primary objective of the study is to investigate the safety, tolerability, and pharmacokinetics of treatment with NNZ-2591 in children and adolescents with PTHS.

All safety analyses will be based on the Safety Population.

Safety will be assessed based on incidence of treatment emergent adverse events (TEAEs), serious adverse events (SAEs), changes in vital signs, changes in laboratory measures, neurological, ophthalmological, and physical exam results, [redacted] and changes in ECG's.

Tolerability will be assessed based on daily dosing logs, requiring dosing adjustments or discontinuations, dosing compliance, and AEs related to subject experience.

Pharmacokinetics will be summarized separately outside of this SAP.

7.1 Dosing Summary

Treatment duration and study duration will be derived as stated in Section 5.3 and summarized using descriptive statistics.

The maximum dose level (8 mg/kg BID, or 12 mg/kg BID) reached by each subject will be summarized according to the Dosing Compliance CRF as a frequency distribution. The number of missed doses as defined in Section 5.3 will be summarized. The number of missed doses will only be defined prior to discontinuation of treatment.

7.2 Adverse Events

Adverse events will be recorded from the time of first dose of IMP until Follow-up (Visit 17). Adverse events summaries will only consider treatment emergent adverse events (TEAEs). TEAEs are defined as those adverse events that occurred after the administration of the first dose of study drug. If it cannot be determined whether the adverse event is treatment emergent due to missing or incomplete (partial) onset dates, dates will be imputed for the purposes of determining treatment emergence as described in Section 5.2.1.

Verbatim terms entered into the clinical database via the EDC system will be mapped to preferred terms and system organ classes using version 24.0 of MedDRA.

Summaries that are displayed by SOC and PT will be ordered by descending order of incidence of SOC and PT within each SOC. Adverse event summaries of the following types will be presented for the Safety Population:

- Overall summary of the TEAEs which contain an overview of each item below.
- Subject count and incidence rate of TEAEs and total number of unique TEAEs by MedDRA system organ class and preferred term.
- Subject count and incidence rate of TEAEs by MedDRA preferred term and greatest intensity. Intensity is defined by three categories: "Mild", "Moderate", and "Severe." At each level of subject summarization, a subject is classified according to the greatest intensity if the subject reported one or more events. Adverse events with missing intensity will be considered severe for this summary.
- Subject count and incidence rate of TEAEs by MedDRA preferred term and closest relationship to study drug (Related/Not Related). At each level of subject summarization, a subject is classified according to the closest relationship if the subject reported one or more events. Adverse events with missing relationship will be considered related for this summary.
- Subject count and incidence rate of Serious TEAEs by MedDRA SOC and PT.
- Subject count and incidence rate of TEAEs by dose level, SOC, and PT.
- Subject count and incidence rate of related TEAEs by dose level, SOC, and PT.
- Subject count and incidence rate of TEAEs leading to withdrawal of study medication by MedDRA SOC and PT.
- Subject count and incidence rate of TEAEs leading to dose modification (defined as dose increased, dose reduced, or dose interrupted on the CRF) by MedDRA SOC and PT.
- Subject count and incidence of TEAEs with an outcome of death by MedDRA SOC and PT.
- [redacted]

A separate listing of related TEAEs will be generated.

7.3 Clinical Laboratory Evaluations

Blood will be collected at Visit 1 (Screening), Visit 3 (Week 0), and Visit 13 (End of Treatment). Urine samples will be collected at all visits. Laboratory parameters (chemistry, hematology, urinalysis) will be presented in data listings. Abnormal values will be flagged as high or low relative to normal ranges, where applicable, by the laboratory vendor.

Change from baseline in quantitative laboratory parameter values will be summarized using descriptive statistics at all visits.

For those laboratory tests in which normal ranges are available, shifts from baseline will be presented. Categories for the shift tables will be high, normal, and low.

Laboratory parameters will be classified according to the categories given by the laboratory vendor. The laboratory parameters from the vendor listed below will be assigned to the following categories for data summaries.

Chemistry	Hematology	Urinalysis
Bun/Creatinine Ratio EGFR (Bedside Schwartz) Hemoglobin A1C Pregnancy, qualitative Thyroxine, Free Triiodothyronine, Free TSH Ultrasensitive Urine Drug Screen w/ Con	PT/INR/APTT	Microalbumin, UR Random Urine Chemistry

7.4 Physical and Neurological Examination [redacted]

The physical examination will be conducted at all in-clinic visits. The exam will note Normal and Abnormal readings. For each abnormal reading, clinical significance will be determined. The physical exam procedure will include the following organ systems: Head, ears, eyes, nose, and throat; Skin; Cardiovascular; Respiratory; Gastrointestinal; Genitourinary (as appropriate or indicated); Musculoskeletal; and Allergies. The exam will include an assessment for sign of bladder or kidney pathology including an exam of the abdomen and both flanks. Assessment of edema will also be done. Exam may include suprapubic exam (bladder palpation) if decreased urine output is noted.

A neurological examination will be conducted at all in-clinic visits. A standard neurological exam will be conducted to include an evaluation of mental status, cranial nerves, motor exam, sensory exam, cerebellar (coordination, gait), and reflexes.

A modified version of the physical and neurological exams will be done by a physician via telemedicine at the remote visits. As part of the in-home visit, the nurse will conduct a targeted physical exam including assessment of general appearance, an abdominal and a costovertebral angle tenderness exam, assessment for signs of cystitis, and an assessment of edema. Exam may include suprapubic exam (bladder palpation) if decreased urine output is noted.

[redacted]

All examination results will be reported as normal or abnormal. Clinical significance will be determined for abnormal results.

The results of all examinations will be presented in data listings.

7.5 Ophthalmology Assessment

A comprehensive ophthalmological exam will be undertaken at Screening (Visit 1) and End of Treatment (Week 13, Visit 16). Routine visual inspection of each eye was undertaken at in-clinic Visit 5 (Week 2) and Visit 9 (Week 6) for subjects enrolled under protocol version 2.0, 2.1 and 2.2. At these interim visit exams, an exam with dilation may be done if abnormalities are present or the assessment is indeterminate, or as deemed necessary by the ophthalmologist, Medical Monitor or Principal Investigator. The complete exam conducted at Visit 1 and Visit 16 could include up to 2 attempts at assessments of estimated visual acuity, dilated refractive error measurement, dilated slit lamp examination of the lens, and a dilated fundus examination.

The results will be presented in data listings.

7.6 Vital Signs

Weight, blood pressure (systolic and diastolic in mmHg), heart rate (bpm), respiratory rate (rpm), and body temperature (°C) will be collected at the Screening visits (Visit 1 and 2) and all post-baseline study visits. Ideally, vital signs will be taken after the subject has been seated for 5 minutes. The initial blood pressure reading will be summarized in the table. Height will be taken at the Screening visit (Visit 1), Visit 3 (Week 0), and End of Treatment (Week 13, Visit 16).

Summary statistics for each measurement and the change from baseline will be presented at each post-baseline visit. All vital sign measurements will be presented in data listings. A separate table and listing will be provided for blood pressure.

7.7 Electrocardiogram (ECG)

The ECG assessment will measure PR Interval (msec), QRS Duration (msec), QT Interval (msec), QTcF Interval (msec), Heart Rate (bpm), and RR Interval (msec). The assessment will be done at the Screening visit (Visit 1), Visit 3 (Week 0), Visit 5 (Week 2), Visit 9 (Week 6), End of Treatment (Week 13, Visit 16) and Follow-Up (Visit 17). The interpretation of the results will be assessed as Normal or Abnormal, and any Abnormal results will be assessed for clinical significance. Ideally, the subject should be supine for at least 5 minutes before the first ECG is taken. If this cannot be done, this will be documented in the CRF. The ECG will be obtained as a continuous ECG of at least 50 ms. All readable ECGs of at least 10 ms will be interpreted by the central reader and reported.

Summary statistics for each ECG parameter and the change from baseline will be presented at each post-baseline visit. Additionally, summary statistics for the shifts in interpretation from baseline will be presented at each post-baseline visit. All ECG measurements will be presented in data listings.

7.8 Prior and Concomitant Medications

Prior and concomitant medication verbatim terms captured via the EDC system will be mapped to Anatomical/Therapeutic/Chemical (ATC) class and Preferred Names using the March 1, 2023 version of the World Health Organization (WHO) Drug Dictionary Enhanced available.

Prior medications are those medications started prior to and stopped before the first dose of study drug. Concomitant medications include: 1) medications started prior to, but still ongoing as of the first dose of study drug, 2) medications with a start date on or after first dose date. Prior and concomitant medications will be summarized separately by WHO ATC class and preferred name. These summaries will present the number and percent of subjects using each medication. Subjects may have more than one medication per ATC class and preferred name. At each level of subject summarization, a subject is counted once if one or more medications at that level is reported for the subject. The drug classification will be the ATC classification 4. If this classification is not present, the next available classification will be utilized. Medications taken with a frequency of PRN or "as needed" will be excluded from the tabular summaries. Each summary will be ordered by descending order of incidence of ATC class and preferred term.

7.9 Monthly Seizure Episodes

Monthly seizure frequency will be derived as defined in Section 5.3. Seizure frequency will be assessed from a seizure log that is part of the Caregiver Diary which is completed daily by the caregiver.

Summary statistics on the observed and the change from baseline in monthly seizures will be displayed at baseline and end of treatment. Within subject changes will be assessed using a Wilcoxon Signed Rank Test.

7.10 Monthly Major Breathing Episodes

Monthly frequency of major breathing events will be derived as defined in Section 5.3. Breathing event frequency will be assessed from a Major Respiratory Events log that is part of the Caregiver Diary which is completed daily by the caregiver.

Summary statistics on the observed and the change from baseline in monthly breathing episodes will be displayed at baseline and end of treatment. Within subject changes will be assessed using a Wilcoxon Signed Rank Test.

8 Efficacy Analysis

The secondary objective is to investigate measures of efficacy during treatment with NNZ-2591 Oral Solution in children and adolescents with PTHS.

All efficacy analyses will be conducted on the ITT and mITT populations.

8.1 Efficacy Variables

There are no primary efficacy variables.

Secondary variables for efficacy to be considered include:

- PTHS-specific Clinical Global Impression Scale–Overall Improvement (CGI-I) Score
- PTHS-specific Clinical Global Impression Scale–Domain Improvement Scores
- Clinical Global Impression Scale - Severity (CGI-S) - Overall Severity Score
- Clinical Global Impression Scale - Severity (CGI-S) - Domain Severity Scores
- MacArthur-Bates Communicative Development Inventory (MB-CDI) Scores
- Observer-Reported Communication Ability (ORCA) Score
- Caregiver Top 3 Concerns Likert Scale Scores

- Aberrant Behavior Checklist-2 (ABC-2) Scores
- Behavior Problems Inventory-Short Form (BPI-SF) Scores
- Child Sleep Habits Questionnaire (CSHQ) Scores
- Gastrointestinal Health Questionnaire (GIHQ) Scores
- Caregiver Impression of Change Scores
- Vineland Adaptive Behavior Scales-3 Scores
- Quality of Life Inventory-Disability (QI-Disability) Scores
- Impact of Childhood Neurological Disability (ICND) Overall quality of life rating
- Modified Two-minute walk test (2MWT) total distance
- SB5 raw, z-deviation, age-equivalent, and Change Sensitive scores or BSID-4 scores (raw, GSV, age-equivalent; standard and scaled scores may be assessed for participants for whom the standard scores can be calculated)

Biomarker variables and gut microbiome analyses will be described in a separate analysis plan.

8.2 Adjustments for Covariates

A list of potential covariates to be used in descriptive summaries or analyses are provided below.

- Age at Baseline
- Sex (Male, Female)
- SB5 NVIQ or BSID-4 NVDQ at Baseline
- History of developmental regression (yes, no)
- ASD diagnosis (yes, no)
- CGI-S Overall Score at Baseline

8.3 Examination of Subgroups

No subgroup analyses are planned.

8.4 Multicenter Studies

All sites will be pooled for this analysis.

8.5 Multiplicity Comparison/Multiplicity

No formal statistical analysis is planned; therefore, there will be no adjustments for multiplicity.

9 Efficacy Analysis Methods

9.1 Primary Efficacy Analysis

There is no primary efficacy endpoint.

9.2 Secondary Efficacy Analysis

The efficacy endpoints will be analyzed using a Wilcoxon signed rank test on the paired subject data from baseline to end of treatment (Visit 16) and Visit 9 (Week 6) when applicable. Descriptive statistics will be used to summarize the data. For total scores (and certain subscale score as defined below) waterfall plots and histograms or bar charts (as appropriate) will also be created at the End of Treatment (EOT) visit with plots for all subjects and separate plots indicating each of the following: age groups, NVIQ/DQ score (>= 50, < 50), sex (Male, Female) and number of concomitant medications (> 6, ≤ 6). The number of concomitant medications will be determined as defined in Section 5.3. Line graphs plotting scores by visit will also be generated for the mean group data and showing individual participants. The NVIQ/DQ score is the SB5 Non-Verbal Intelligence Quotient (z-deviation), or the BSID-4 Non-verbal Developmental Quotient depending on which assessment the subject took. The number and percentage of subjects who had an IQ/DQ score >= 50 and < 50 at Baseline and End of Treatment will be presented in a table for NVIQ/DQ, VIQ/DQ, and FSIQ/Overall DQ. For assessments that use discrete scores, the histograms will be changed to bar charts. If the change from baseline of the discrete scores is presented, then the change from baseline value will be rounded using conventional rounding standards for ease of presentation and understanding in the bar chart. For scales with domain scores, forest plots displaying the mean change from baseline and 95% CI for each domain score at EOT will be generated. All figures will be generated on the ITT and mITT populations.

Exploratory analyses on some endpoints using linear regression may be performed to assess the relationship between treatment and specified covariates or combinations of covariates. The empirical cumulative distribution function will be plotted for certain variables. As appropriate, collective response analysis may be done to assess data across subjects.

A correlation analysis of the efficacy measures will be done using Pearson correlation coefficients at Baseline and EOT. The correlation of each total score will be measured against the others and presented in a table and a heat map/matrix. CGI-I and CGI-S domain scores will also be included in the correlation analysis. Test-retest reliability will be assessed for continuous efficacy assessments collected more than once during the Screening period (Visits 1, 2, and 3, pre-dosing) using Intraclass Correlation Coefficients

(ICC). The ICC will be based on covariance parameter estimates from a Restricted Maximum Likelihood Mixed Model including results from assessments conducted during the Screening/Baseline period with a random effect for subject. The ICC will be derived as: ICC = Within-subject covariance/Total covariance. ICC values range from 0 to 1, with values closer to 1 indicating greater test-retest reliability.

A treatment policy estimand will be used for all endpoints. The variables of interest are numeric scores associated with subject responses. Summaries will be performed on the observed data. There will be no imputations for completely missing scores or intercurrent events. Handling of missing responses items that contribute to total and subscale scores are defined in Section 5.2.5.

Results for each efficacy variable will also be listed by subject.

9.2.1 Clinical Global Impression Scale - Overall Improvement Score

CGI-I assessments are completed at Visit 9 (Week 6), End of Treatment (Week 13, Visit 16), and Follow-up (Visit 17). The clinician rates how much the subject’s illness has improved or worsened relative to Visit 3 (Week 0). The Overall CGI-I score is rated on a 7-point scale, where 1=’very much improved’; 2=’much improved’; 3=’minimally improved’; 4=’no change’; 5=’minimally worse’; 6=’much worse’; 7=’very much worse’. Scores of 1, 2, and 3 indicate improvement. Within subject change is indicated by the score at each relevant follow-up visit (Week 6, Visit 9 and Week 13, EOT, Visit 16). If an overall score is missing, that subject will be excluded from analysis at that visit.

Summary statistics for the CGI-I Overall score will be displayed at all post-baseline visits. Within subject changes will be analyzed using a Wilcoxon signed rank test at EOT and Week 6. The overall score will be presented in figures as described in Section 9.2. As this score is categorical, the waterfall plots will be adjusted so a score of “7 – Very Much Worse” will be presented at the bottom of the graph and a score of “1 – Very Much Improved” will be presented at the top of the graph. The overall score will also be presented as a line graph at post-baseline visits.

9.2.2 Clinical Global Impression Scale - Domain Improvement Scores

CGI-I assessments are completed at Visit 9 (Week 6), End of Treatment (Week 13, Visit 16), and Follow-up (Visit 17). The clinician rates how much the subject’s illness has improved or worsened relative to Visit 3 (Week 0). The seven domain improvement scores are: Language/Communication, Social Interaction, Ambulation/Gross Motor, Fine Motor Functioning/Self-Help, Gastrointestinal, Autonomic/Breathing Abnormalities, and Challenging Behaviors: Repetitive Behavior/Self-Injury/Aggression. The individual domain CGI-I scores are rated on a 7-point scale, where 1=’very much improved’;

2=’much improved’; 3=’minimally improved’; 4=’no change’; 5=’minimally worse’; 6=’much worse’; 7=’very much worse’. Scores of 1, 2, and 3 indicate improvement. If a domain score is missing, that subject will be excluded from analysis for that domain.

Within subject change is indicated by the score at each relevant follow-up visit (Week 6, Visit 9 and Week 13, EOT, Visit 16). If a domain score is missing, that subject will be excluded from analysis for that domain at the relevant visit. The domain scores will be presented in figures as described in Section 9.2. Domain scores at EOT will be presented in a forest plot. Domain scores will be presented as line graphs at post-baseline visits.

Domain scores as reported on the CRF will be summarized at each post-baseline visit, in the same manner as the CGI-I Overall Score.

9.2.3 Caregiver Impression of Change Score

The Caregiver Impression of Change assessment will be completed at the end of treatment visit (Week 13, Visit 16).

The caregiver is asked to rate the change in his or her child’s overall function and well-being since Visit 3 (Week 0), and the change in specific symptoms. This is rated on a 7-point Likert scale: 1-Very much improved, 2-Much improved, 3-Improved, 4-Unchanged, 5-Worse, 6-Much worse, 7-Very much worse. Scores of 1, 2 or 3 indicate improvement in overall function and symptoms.

The assessment of change should be made relative to Visit 3 (Week 0). The caregiver should assess the subject’s current state at the visit taking into consideration observations at the visit and symptoms within the last week.

For the rating of overall function, the caregiver also identifies the one symptom area or feature that has most influenced his or her rating of the child’s overall function and provides a description of symptom changes and impact of the changes.

The symptom domains rated are: communication, social interaction, behavior, motor abilities, seizures, cognition and learning, self-care, gastrointestinal (GI) problems, and breathing problems. Within subject change is indicated by the score at EOT. If a score is missing for a domain, that subject will be excluded from the corresponding domain summaries.

Summary statistics on the Caregiver Impression of Change overall and domain scores will be displayed at End of Treatment. The overall score will be presented in waterfall plots, and histograms or bar charts (as appropriate), and the domain scores will be presented in a forest plot, as described in Section 9.2.

9.2.4 Clinical Global Impression Scale - Severity - Overall Severity Score

Clinical Global Impression of Severity assessments will be made at the Screening visit (Visit 1), Visit 3 (Week 0), Visit 9 (Week 6), End of Treatment (Week 13, Visit 16), and Follow-up (Week 15, Visit 17). The overall severity score is scored on a 7-point scale, where 1='typical for age, not at all impaired'; 2='borderline, slightly impaired'; 3='mildly impaired'; 4='moderately impaired'; 5='markedly impaired'; 6='severely impaired'; 7='among the most severely impaired'. Lower scores reflect less impairment. Within subject change is the difference between the score at each relevant follow-up visit and the score at baseline. If a score is missing, the subject will be excluded from the corresponding summaries.

The overall score and change from baseline at each post-baseline visit will be summarized using descriptive statistics. Within subject changes will be analyzed using a Wilcoxon signed rank test at Week 6 and EOT. Test-retest reliability will be assessed for all scores collected during the Screening period (Visits 1 and 3) using ICC. The change from baseline in the overall score at EOT will be presented in figures as described in Section 9.2. The overall score will also be presented as a line graph at post-baseline visits.

9.2.5 Clinical Global Impression Scale - Severity - Domain Severity Scores

The CGI-S assessments will be made at the Screening visit (Visit 1), Visit 3 (Week 0), Visit 9 (Week 6), End of Treatment (Week 13, Visit 16), and Follow-up (Week 15, Visit 17). The seven domain severity scores are: Language/Communication, Social Interaction, Ambulation/Gross Motor, Fine Motor/Self-Help, Gastrointestinal, Autonomic/Breathing Abnormalities, and Challenging Behaviors: Repetitive Behavior/Self-Injury/Aggression. Each domain score is scored on a 7-point scale, where 1='typical for age, not at all impaired'; 2='borderline, slightly impaired'; 3='mildly impaired'; 4='moderately impaired'; 5='markedly impaired'; 6='severely impaired'; 7='among the most severely impaired'. Lower scores reflect less impairment. Within subject change is the difference between the total score at each relevant follow-up visit and the score at baseline. If a score is missing, the subject will be excluded from the corresponding summaries.

The change from baseline in CGI-S domain scores will be analyzed in the same manner as the CGI-S overall score. The change from baseline in domain scores at EOT will be presented in forest plots as described in Section 9.2. Domain scores will be presented as line graphs at post-baseline visits.

9.2.6 MacArthur-Bates Communicative Development Inventory Scores

The MB-CDI will be completed at Visit 3 (Week 0) and End of Treatment (Week 13, Visit 16). This caregiver-completed instrument assesses language development in vocabulary comprehension and production, gestures, and grammar. Designed to assess early language development in infants and toddlers, it can also be used with older, developmentally delayed children to assess developmental progress in language development. The MB-CDI includes three modules based on chronological or developmental age: Words and Gestures (8 to 18 months or developmental equivalent), Words and Sentences (16 to 30 months or developmental equivalent) and the CDI-III, an extension of the MB-CDI used with children 30 to 37 months (or developmental equivalent). Only the Words and Gestures module will be used for this study in line with the cognitive and language development of the patient population.

The number of “yes” answers in “First signs of understanding” and “sometimes” and “often” in “Starting to Talk” will be summarized. The number of phrases understood, the number of words understood, number of words produced, number of early gestures, number of later gestures, and number of gestures will be summarized. For these scores, scores calculated by the site rater and entered in the EDC will be used for the analysis. Scores may also be derived if the EDC score calculations were not completed. The calculations are as follows:

- Phrases understood: total count of phrases selected in “Phrases” section (total phrases 28)
- Words Understood: total count of words selected as either “understands” or “understands and says” in the “Vocabulary Checklist”. If for any given word both “understands” and “understands and says” are selected, the word is counted only once. Blanks are not counted. Total words 396.
- Words Produced: total count of words selected as “understands and says”. If for any given word both “understands” and “understands and says” are selected, the word is counted only once. Blanks are not counted. Total words 396.
- Early Gestures: total count of “sometimes” and “often” responses in “First Communicative Gestures” and “yes” answers in “Games and Routines”. 18 maximum.
- Later Gestures: total count of “yes” answers in “Actions with Objects,” “Pretending to be a Parent,” and “Imitating Other Adult Actions”. 45 maximum.
- Total Gestures: Total of Early Gestures and Later Gesture scores. 63 maximum.

Change from baseline for the scores will be analyzed using a Wilcoxon signed rank test. The change from baseline in the “words understood” score and the “words produced score” at EOT will be presented in plots as described in Section 9.2. Exemplars of new words may be listed for individual subjects who show increases in vocabulary from baseline based on an increase in number of words produced.

9.2.7 Observer-Reported Communication Ability Score

The ORCA measure will be completed at Visit 3 (Week 0) and End of Treatment (Week 13, Visit 16). This caregiver-reported measure was developed to assess communicative abilities in verbal and low-verbal/non-verbal populations. The caregiver evaluates the individual's observed communication ability over the past 30 days. The ORCA measure will be scored using the R program provided by the ORCA developers (Center for Health Measurement, Duke University). The macro will produce an ORCA t-score as long as at least one of the 20 concepts or items has a score. A concept score will be given only if all items within the concept are answered. Higher scores reflect greater communication ability. The ORCA t-score range is 25.8 to 83.8.

Alterative scoring algorithms are in the process of being developed by the ORCA developers based on feedback from users in the research community. These alternative algorithms would provide scoring alternatives to provide credit for emerging skills in addition to mastery of skills. These algorithms are in development and are not yet available from the ORCA developers. In line with these approaches, a modified t-score will also be calculated as an exploratory measure. The SAS macro used to calculate the total score will be edited to give credit for both "always" and "sometimes" responses (to assess emerging skills). This modification to the analysis of the items will be used to calculate the modified t-score.

The t-score and modified t-score will be summarized using descriptive statistics. Change from baseline will be analyzed using a Wilcoxon signed rank test. The change from baseline in the t-score and modified t-score at EOT will be presented in plots as described in Section 9.2.

9.2.8 Caregiver Top 3 Concerns Rating

The Caregiver Top 3 Concerns Rating assessment will be completed at all Screening/Baseline period visits (Visits 1, 2, 3), as well as Week 6 (Visit 9), End of Treatment (Week 13, Visit 16), and Follow-up (Visit 17). The three concerns identified at Screening/Baseline are carried forward and rated at the follow-up visits.

Caregivers are asked to identify 3 concerns from among 11 clinically important symptom domains: Language/Communication, Social interest/Avoidance, Repetitive Behaviors, Challenging Behaviors, Gross motor skills, Fine motor skills, Gastrointestinal problems, Breathing Disruptions, Sleep, Seizures, and Self-care skills. Each individual concern identified should be from a different symptom domain, but caregivers can select any domain and choose any symptom of concern within that domain.

The severity of each concern is scored using a 10-point Likert scale of severity. Average Top 3 concern severity is defined as the average of severity scores for the three concerns. Lower scores indicate lesser severity. The average score (i.e., average of the severity scores for the Top 3 concerns) will have a range of 0 to 10. Improvement in concern average severity is defined as a change from baseline score < 0.

If the caregiver does not list the same 3 concerns on the Top 3 Concerns Follow-Up CRF that were recorded on the Top 3 Concerns Screening CRF (completed at Visit 1), then their responses will be reviewed and may not be included in analysis. A review of all concerns will be done prior to database lock, and visits with incorrect concerns will be flagged as unusable for analysis.

The frequency of any symptom being chosen for all concerns combined will be summarized. Summary statistics for the change from baseline in the average severity and by symptom domain will be displayed at all post-baseline visits. Change from baseline will be analyzed using a Wilcoxon signed rank test at Week 6 and EOT. The frequency of subjects with improvement in the concern severity will be summarized by average concern and symptom domain. Test-retest reliability will be assessed for all scores collected during the Screening period (Visits 1, 2, and 3, pre-dosing) using ICC. The change from baseline in the average concern severity at EOT will be presented in plots, as described in Section 9.2. The change from baseline in domain severity scores at EOT will be presented in a forest plot. The domain and average concern severity will be presented in line graphs at each visit.

9.2.9 Aberrant Behavior Checklist-2 Scores

The ABC-2 will be completed at all Screening/Baseline period visits (Visits 1, 2, 3), as well as Week 6 (Visit 9), End of Treatment (Week 13, Visit 16), and Follow-up (Visit 17). Each item is scored on a 4-point scale, where 0= Not at all a problem; 1= The behavior is a problem but slight in degree; 2= The problem is moderately serious; 3= The problem is severe in degree. The items are grouped into 5 subscales: Irritability (Subscale I, score ranges from 0 – 45), Social Withdrawal (Subscale II, score ranges from 0 – 48), Stereotypic Behavior (Subscale III, score ranges from 0 – 21), Hyperactivity/Noncompliance (Subscale IV, score ranges from 0 – 48), and Inappropriate Speech (Subscale V, score ranges from 0 – 12). The total score has a range from 0 – 174. Lower scores indicate fewer behavior issues. The subscale scores are calculated via a scoring algorithm in the EDC. These calculated scores will be used for the analysis. If an item is missing and as such a subscale and total score was not calculated in the EDC, the missing items will be imputed and subscale or total scores calculated from answered and imputed items. Missing items will be imputed as defined in section 5.2.5. The total score will only be calculated if all items are answered or can be imputed. If a total score is missing after applying the imputation algorithm, then that subject will be excluded from the corresponding summaries. The subscales and their corresponding items are listed below.

- Subscale I (Irritability): Items 2, 4, 8, 10, 14, 19, 25, 29, 34, 36, 41, 47, 50, 52, 57.
- Subscale II (Social Withdrawal): Items 3, 5, 12, 16, 20, 23, 26, 30, 32, 37, 40, 42, 43, 53, 55, 58
- Subscale III (Stereotypic Behavior): Items 6, 11, 17, 27, 35, 45, 49
- Subscale IV (Hyperactivity/Noncompliance): Items 1, 7, 13, 15, 18, 21, 24, 28, 31, 38, 39, 44, 48, 51, 54, 56
- Subscale V (Inappropriate Speech): Items 9, 22, 33, 46

The subscale scores are calculated as the sum of item scores within the subscale. The total score is calculated as the sum of all items across all subscales.

Within subject change is the difference between the score at each relevant follow-up visit and the score at baseline. Summary statistics for the change from baseline in total and each subscale ABC-2 score will be displayed at all post-baseline visits. Change from baseline in the total and subscale scores will be analyzed using the Wilcoxon signed rank test at Week 6 and EOT. Test-retest reliability will be assessed for all scores collected during the Screening period (visits 1, 2, and 3, pre-dosing) using ICC. The change from baseline in the total score at EOT will be presented in plots as described in Section 9.2. The change from baseline in subscale scores at EOT will be presented in a forest plot. The total and subscale scores will be present in line graphs at each visit.

9.2.10 Behavior Problems Inventory-Short Form

The BPI-SF will be completed at Visit 3 (Week 0) and at the End of Treatment (Week 13, Visit 16). The BPI-SF is a caregiver-completed rating scale for assessing self-injury, aggressive and stereotyped behavior. Each item is measured for frequency and items in the Self-Injurious Behavior and Aggressive/Destructive Behavior subscales are also assessed for severity. The frequency is scored on a 5-point scale, where 0=never/no problem; 1=monthly, 2=weekly, 3=daily, 4=hourly. The severity is scored on a 3-point scale, where 1=mild, 2=moderate, 3=severe. The items are grouped in three subscales: Self-Injurious Behavior (score ranges from 0 – 32 for frequency and 8 – 24 for severity), Aggressive/Destructive Behavior (score ranges from 0 – 40 for frequency and 10 – 30 for severity), and Stereotyped Behavior (score ranges from 0 – 48 for frequency). The score for each subscale is the sum of the items within that domain. The total frequency score is the sum of the frequency scores for the three domains. The total severity score is the sum of the severity scores for Aggressive/Destructive Behavior and Self-Injurious Behavior subscales. Missing items will be imputed as defined in Section 5.2.5. A subscale score will only be calculated if all items in that domain are answered or can be imputed. Lower scores indicate lesser frequency and severity in behavior problems. Within subject change is the difference between the score at EOT and the score at baseline.

Summary statistics for the total score, subscale scores and their changes from baseline in the frequency and severity of BPI-SF scores will be displayed at End of Treatment. Within subject changes will be analyzed using a Wilcoxon signed rank test. The change from baseline in the total frequency and severity scores at EOT will be presented in plots as described in Section 9.2. The change from baseline in the subscale scores at EOT will be presented in a forest plot and waterfall plots. The subscale and total scores will be presented in line graphs at each visit.

9.2.11 Child Sleep Habits Questionnaire Scores

The CSHQ will be completed at all Screening/Baseline period visits (Visits 1, 2, 3), as well as Week 6 (Visit 9), End of Treatment (Week 13, Visit 16), and Follow-up (Visit 17). Each item is scored on a 3-point scale, where 1=rarely; 2=sometimes; 3=usually. Items 32 and 33 are scores as 1 = not sleepy, 2 = very sleepy and 3 = falls asleep. A higher score reflects more disturbed sleep behavior. The items are grouped into 4 domains and 8 subscale scores are derived from the items in these domains. The subscale scores are a sum of the items within that subscale, except for a few items which are reversed for scoring. Missing items will be imputed as defined in Section 5.2.5. The subscale scores will only be calculated if all items in the subscale are answered or can be imputed. The subscales and their corresponding items are listed below.

- Bedtime Resistance: Items 1, 3, 4, 5, 6, 8, where items 1 and 3 are reversed. Subscale score ranges from 6 – 18.
- Sleep Onset Delay: Item 2 where item 2 is reversed. Subscale score ranges from 1 – 3.
- Sleep Duration: Items 9, 10, 11, where items 10 and 11 are reversed. Subscale score ranges from 3 – 9.
- Sleep Anxiety: Items 5, 7, 8, 21. Subscale score ranges from 4-12.
- Night Wakings: Items 16, 24, 25. Subscale score ranges from 3-9.
- Parasomnias: Items 12, 13, 14, 15, 17, 22, 23. Subscale score ranges from 7-21.
- Sleep Disordered Breathing: Items 18, 19, 20. Subscale score ranges from 3-9.
- Daytime Sleepiness: Items 26, 27, 28, 29, 30, 31, 32, 33, where item 26 is reversed. Subscale score ranges from 8-24.

The total score will be calculated as a sum of all 33 items, with items 1 – 3, 10 – 11, and 26 reversed. The total score will range from 33 – 99 and will only be calculated if all items are answered or can be imputed. Missing items will be imputed as defined in Section 5.2.5. Within subject change is the difference between the score at each relevant follow-up visit and the score at baseline.

Summary statistics for the scores and their changes from baseline in the total score and subscale CSHQ scores will be displayed at all post-baseline visits. Test-retest reliability will be assessed for all scores collected during the Screening period (Visit 1, 2, and 3, pre-dosing) using Intraclass Correlation Coefficients (ICC). Changes from baseline will be analyzed using a Wilcoxon signed rank test at Week 6 and EOT. The change from baseline in the total score at EOT will be presented in plots as described in Section 9.2. The change from baseline in the subscale scores at EOT will be presented in a forest plot, as described in Section 9.2. The subscale and total scores will be presented in line graphs at each visit.

9.2.12 Gastrointestinal Health Questionnaire Scores

The GIHQ will be completed at all Screening/Baseline period visits (Visits 1, 2, 3), as well as Week 6 (Visit 9), End of Treatment (Week 13, Visit 16), and Follow-up (Visit 17). Each item is scored on its frequency, relevancy, and importance to the caregiver. The Frequency is measured on a 5-point scale where 0=never; 1=almost never; 2=sometimes; 3=often; 4=almost always. The surgery domain records occurrence as 0=No and 1=Yes. The relevancy is scored on a scale from 1 to 4, where 1 is not relevant and 4 is very relevant. The importance is scored on a scale from 1 to 4, where 1 is not important and 4 is very important. Lower frequency scores indicate fewer gastrointestinal problems. For relevancy, higher scores indicate the question is more relevant and for importance, higher scores indicate the question is more important. The total frequency score is the sum of all item frequency scores. Similarly, the total relevancy and total importance scores are the sum of all item relevancy and importance scores, respectively. The items are grouped into 9 subscales. The subscale score is the sum of all items within the subscale, as shown on the CRF. Missing items will be imputed as defined in Section 5.2.5. The subscale scores will only be calculated if all items within the domain are answered or can be imputed. The subscales are as follows:

- General Health/Pain: Items 1 – 5, frequency score ranges from 0 – 20, relevancy score ranges from 5 – 20, importance score ranges from 5 – 20.
- Eating, Chewing, and Swallowing: Items 1 – 9, frequency score ranges from 0 – 36, relevancy score ranges from 9 – 36, importance score ranges from 9 – 36.
- Reflux: Items 1 – 3, frequency score ranges from 0 – 12, relevancy score ranges from 3 – 12, importance score ranges from 3 – 12.

- Gas and Bloating: Items 1 – 5, frequency score ranges from 0 – 20, relevancy score ranges from 5 – 20, importance score ranges from 5 – 20.
- Diarrhea and Constipation: Items 1 – 6, frequency score ranges from 0 – 24, relevancy score ranges from 6 – 24, importance score ranges from 6 – 24.
- Personality and Mood: Items 1 – 5, frequency score ranges from 0 – 20, Relevancy score ranges from 5 – 20, importance score ranges from 5 – 20.
- Medications: Items 1 – 9, frequency score ranges from 0 – 36, relevancy score ranges from 9 – 36, importance score ranges from 9 – 36.
- Surgery: Items 1 – 5, frequency score ranges from 0 – 5, relevancy score ranges from 5 – 20, importance score ranges from 5 – 20.
- Parenting: Items 1 – 8, frequency score ranges from 0 – 24, relevancy score ranges from 8 – 24, importance score ranges from 8 – 24.

Within subject change is the difference between the score at each relevant follow-up visit and the score at baseline. Summary statistics for the change from baseline in the total and each GIHQ domain score will be displayed at all post-baseline visits. Test-retest reliability will be assessed for all scores collected during the Screening period (Visits 1, 2, and 3, pre-dosing) using ICC. Changes from baseline will be analyzed using a Wilcoxon signed rank test at Week 6 and EOT. The change from baseline in the total frequency score at EOT will be presented in plots as described in Section 9.2. The change from baseline in the domain frequency scores at EOT will be presented in a forest plot and waterfall plots. The domain and total frequency scores will be presented in line graphs at each visit, with the exception of the Surgery domain, which is scored differently from the other domains and is not comparable.

9.2.13 Vineland Adaptive Behavior Scales-3 (VABS-3)

The VABS-3 (Comprehensive Interview version) will be completed at Visit 3 (Week 0) and End of Treatment (Week 13, Visit 16) as a clinical interview. Scores are calculated by the site rater. An Adaptive Behavior Composite (ABC) score will be calculated, along with a standard score for each of the 4 domains: Communication, Daily Living Skills, Socialization, Motor Skills. For the motor skills domain, standard scores can only be calculated for participants ages 3-9. These range from 20 to 140.

In each of the 11 subdomains, the raw scale (range 0 to 116), v-scale (range 1 to 24), age equivalent (range 0:0 to 22:0+), and growth scale (range 10 to 197) scores are recorded. Lower scores reflect less functional abilities. If a score is missing, that subject will be excluded from corresponding summaries. Within subject change is the difference between the score at EOT and the score at baseline.

Summary statistics for the change from baseline in the Vineland Adaptive Behavior Scales-3 composite and domain standard scores, and subdomain raw, v-scale, and growth scale scores will be displayed at end of treatment. Changes from baseline will be analyzed using a Wilcoxon signed rank test. The change from baseline in the composite standard score, raw scores and GSV at EOT will be presented in plots as described in Section 9.2. The change from baseline in the domain standard scores and subdomains at EOT will be presented in a forest plot. The domain standard scores and subdomain scores will be presented in line graphs at each visit.

9.2.14 Quality of Life Inventory - Disability (QI-Disability) Scores

The QI-Disability Score will be completed at Visit 3 (Week 0) and at the End of Treatment (Week 13, Visit 16). Each item is scored on a 5-point scale, where 1=never; 2=rarely; 3=sometimes; 4=often; 5=very often. The items are grouped into six domains for analysis: physical health (items 1-4), positive emotions (items 5-8), negative emotions (items 9-15), social interaction (items 16-22), leisure and the outdoors (items 23-27), and independence (items 28-32). The domain scores are calculated by first transforming the 5-point scale, so that 1=0, 2=25, 3=50, 4=75, and 5=100. Items 9-15 are reversed in scoring, so that 1=100, 2=75, 3=50, 4=25, and 5=0. The domain score is the average of the transformed score of the items within that domain. Missing items will be imputed as defined in Section 5.2.5. A domain score will be calculated as long as at least one of the items within the domain is answered. The overall score is the average of all domain scores and will be calculated as long as at least one of the domain scores can be calculated. Higher scores indicate a better quality of life. Within subject change is the difference between the score at EOT and the score at baseline.

Summary statistics for the change from baseline in the QI-Disability overall and domain scores will be displayed at end of treatment. Changes from baseline will be analyzed using a Wilcoxon signed rank test. The change from baseline in the overall score at EOT will be presented in plots as described in Section 9.2. The change from baseline in the domain scores at EOT will be presented in a forest plot. The domain and overall scores will be presented in line graphs at each visit.

9.2.15 Impact of Childhood Neurological Disability Overall quality of life rating

The ICND Scale will be administered at Visit 3 (Week 0) and at End of Treatment (Week 13, Visit 16). The ICND is scored on a scale from 1 to 6, where 1=poor and 6=excellent. Higher scores indicate better quality of life. Within subject change is the difference between the score at EOT and the score at baseline. If a score is missing, that subject will be excluded from corresponding summaries. If in the EDC, the numeric score is missing and only the descriptor is present, the numeric equivalent score will be used for analysis.

Summary statistics for the observed and change from baseline in the ICND score will be displayed at end of treatment. Change from baseline will be analyzed using a Wilcoxon signed rank test. The change from baseline in the ICND score at EOT will be presented in plots, as described in Section 9.2.

9.2.16 Bayley Scales of Infant and Toddler Development version 4

The BSID-4 will be performed at Screening (Visit 1) and End of Treatment (Week 13, Visit 16) on those subjects who cannot be assessed using the SB5. It assesses cognitive ability, motor skills, and behavior and provides standardized scores in five domains: Cognitive (COG), Language (LANG), Motor (MOT), Social-Emotional (SOEM), and Adaptive Behavior. The Adaptive Behavior scale was not collected for this study. Domain scores range from 40 – 160, where higher scores indicate greater ability in the domain area of development. Scaled scores, age equivalents and growth scale values (GSVs) can also be calculated for the scale subtests: cognition (CG), receptive communication (RC), expressive communication (EC), fine motor (FM), and gross motor (GM). Scaled scores and age equivalents can be calculated for the social emotional (SE) subtest. Scale subtest scores range from 1 – 19 with a mean of 10 and SD of 3. Age equivalent scores range from < 0:15 – >41:08, Growth scale scores range from 414 – 559 with a mean of 500 and SD of 25. Higher scores indicate greater development in the scale subtest. Raw scores, age equivalent scores and GSV are collected for all subjects. Standard scores and scaled scores are only calculated for participants 42 months and younger.

As part of the analysis, an overall developmental quotient, a verbal developmental quotient, and a non-verbal developmental quotient will also be calculated. These developmental quotients will only be calculated if the contributing scale scores are all non-missing. The developmental quotients will be derived as follows:

Overall Developmental Quotient:

(Mean of the age-equivalent scores in months of the cognition, fine motor, receptive communication, expressive communications scales/chronological age in months) x 100

Verbal Developmental Quotient:

(Mean of the age-equivalent scores in months of the expressive communication and receptive scales/chronological age in months) x 100

Non-verbal Developmental Quotient:

(Mean of the age-equivalent scores in months of the cognition and fine motor scales/chronological age in months) x 100

Subtest/Subdomain Developmental Quotients:

The developmental quotients for the BSID subtests may also be derived if EDC calculations were not completed. The developmental quotients for each subtest are calculated in a similar manner as the overall, non-verbal, and verbal developmental quotients. Subtest Developmental Quotient = (Age Equivalent Score in Months for the subtest / chronological age in months) x 100.

Chronological age in months will be calculated as (date of assessment – date of birth + 1)/30.4375.

For age equivalent values < 2, the result is entered into EDC as 1.9, for analysis. For age equivalent values > X, where X is a numeric value, the result is entered into EDC as X.1, for analysis. For example, “>33” is entered in EDC as 33.1.

The standard scores have a mean of 100 and a standard deviation of 15. The developmental quotient scores are a ratio of developmental age to chronological age and will have a range of 1 to 100. Higher scores indicate a higher IQ or better developmental progress. If any scores are missing, the subject will be excluded from corresponding summaries.

Within subject change is the difference between the score at EOT and the score at baseline. If a score is missing, that subject will be excluded from the corresponding summaries.

Summary statistics for the observed and change from baseline in the age equivalent, growth scale subtest scores, the standard and scale scores (for participants for whom the scores are calculated), and developmental quotients will be displayed at end of treatment. Changes from baseline will be analyzed using a Wilcoxon signed rank test. The change from baseline in the raw scores and GSV at EOT will be presented in plots as described in Section 9.2. The change from baseline in the raw and GSV domain scores will be presented in a forest plot, as described in Section 9.2.

9.2.17 Modified two-minute walk test

The modified 2MWT be administered at Screening (Visit 1), Visit 3 (Week 0) and at End of Treatment (Week 13, Visit 16). The walk test measures how far an individual can walk during a 2-minute period and assesses endurance which has relevance for everyday functioning. The distance is measured in meters and greater total distance indicates greater physical ability. Within subject change is the difference between the score at EOT and the score at baseline. If a score is missing, that subject will be excluded from corresponding summaries.

Summary statistics for the observed and change from baseline in the total distance measured during the walk test will be displayed at end of treatment. Change from baseline will be analyzed using a Wilcoxon signed rank test. Test-retest reliability will be assessed for all scores collected during Screening period (Visits 1 and 3) using ICC. The change from baseline in total distance at EOT will be presented in plots, as described in Section 9.2.

9.2.18 Stanford-Binet Intelligence Scales version 5

The SB5 will be performed at Screening and Visit 16 (End of Treatment). Participants who cannot be assessed on the SB5 due to their developmental level will be assessed using the BSID-4 instead. General cognitive ability is reported as a full scale intellectual quotient score (FSIQ). Cognitive ability in five categories is reported in the index scores: fluid reasoning (FR), knowledge (KN), quantitative reasoning (QR), visual-spatial processing (VS), and working memory (WM). Non-verbal (NVIQ) and verbal intelligence quotients (VIQ) are also recorded. Scores range from 40 to 160, where higher scores indicate a higher IQ. For the analysis of index scores and IQ, the z-deviation method for intellectual disability will be used and z-deviation scores will be derived based on age at assessment according to the published algorithm (Sansone et al. 2014). Standard scores, change-sensitive scores, age equivalents and the raw scores are calculated and entered by the site rater.

For age equivalent values < 2, the result is entered into EDC as 1.9, for analysis. For age equivalent values > x, where X is a numeric value, the result is entered into EDC as X.1, for analysis. For example, ">33" is entered in EDC as 33.1.

Summary statistics on the observed and change from baseline in the z-deviation scores, change-sensitive scores and age equivalents for FSIQ, NVIQ, VIQ, and the domain scores: FR, KN, QR, VS, WM. Each category will be displayed at end of treatment. The change from baseline in FSIQ, NVIQ, and VIQ z-deviation scores at EOT will be presented in plots as described in Section 9.2. The change from baseline in the domain z-deviation scores will be presented in a forest plot, as described in Section 9.2.

9.3 Exploratory Efficacy Analysis

9.3.1 Biomarkers and gut microbiome

Blood and stool samples will be collected at Screening (Visit 1), Week 0 (Visit 3, before dosing), and at End of Treatment (Visit 8), or upon early termination. Biomarker and gut microbiome will be analyzed separately outside of this analysis plan.

9.3.2 Correlation of Efficacy Endpoints

The potential covariates in this study are listed in Section 8.2. A 2x2 correlation analysis will be performed on all combinations to evaluate the relationship between the covariates. The Pearson Correlation Coefficient will be used to measure the association between age as a continuous variable and the SB5 NVIQ or BSID-4 NVDQ as described in Section 9.2. If the coefficient is <.5, we will assume little to no association. The Point-Biserial Correlation Coefficient will also be used to measure the association between combinations of continuous and binary variables such as SB5 NVIQ/BSID-4 NVDQ and ASD diagnosis, respectively. The binary variables will be assigned a numeric value of 1 or 2. The Pearson Chi-Square test will be used to measure the association between combinations of categorical variables such as age group and genotype or sex and history of regression. A p-value < 0.10 would indicate association between the variables.

The correlation analysis will support the selection of covariates (or combinations thereof) to be used in a linear regression analysis on all pre-specified endpoints. Endpoints of interest will be:

- CGI-I Overall Score
- CGI-S Overall Score
- ORCA t-score
- BPI-SF Total score
- ABC-2 Total Score
- CSHQ Total Score
- GIHQ Total Score
- Vineland ABC score
- Vineland GSV, raw, and standard domain scores
- QI-Disability Overall Score
- ICND Score

9.3.3 Additional exploratory analyses

Additional analyses will be performed to examine the empirical cumulative distribution function of the change from baseline at EOT for the ORCA t-score, CGI-I, CGI-S, ABC-2 total score, CSHQ total score, GIHQ total frequency score, the BPI-SF total score, each BPI-SF domain score, VABS-3 composite score, VABS-3 GSV scores, VABS-3 raw scores, QL-Disability overall score, and ICND score will be plotted.

Test-retest reliability will be assessed for assessments collected more than once during the Screening period (Visits 1, 2, and 3, pre-dosing) using ICC. A correlation analysis of the efficacy measures will be done using Pearson correlation coefficients. The correlation of each total or overall score and CGI-I and CGI-S domain scores at EOT will be measured against the others and presented in a table and a heat map/matrix.

Waterfall plots may also be created for each post-baseline visit with the plots cross-hatched to indicate genotype, History of Regression (yes, no), ASD diagnosis (yes no), depending on the number of subjects in each subgroup.

10 Tables, Listings, and Figures

A list of the tables, listings, and figures will be maintained outside of this document and may be amended as needed.

Data listings will be sorted by subject number, then chronologically (if appropriate) as indicated by visit, date, and time. All data recorded on the CRFs will be included in the listings as appropriate as well as some derived information. Data on other treatments (e.g. behavioral treatments, supplements and special diets) were originally collected on three separate CRFs in the EDC: “Other Treatments”, “Supplements and Complementary Treatments” and “Behavioral/Educational Treatments”. Since the CRFs were duplicative the “Supplements and Complementary Treatments” and “Behavioral/Educational Treatments” were closed and only the “Other Treatments” CRF page was used. For the listing, the data from these three CRF pages will be combined and included in a single listing “Other Treatments.” The “New Symptoms Log” was closed so the listing will only include data up until the time the log was closed in the EDC.

Listings of Subject Disposition, Protocol Deviations, and Inclusion/Exclusion Criteria Not Met will display results for All Subjects. Other listings will display results for the Safety or Intent to Treat Populations as appropriate.

Rounding for all variables will occur only as the last step, immediately prior to presentation in listings, tables, and figures. No intermediate rounding will be performed on derived variables. The standard rounding practice of rounding numbers ending in 0-4 down and numbers ending in 5-9 up will be employed.

Tables will generally be presented with the parameter in the first column followed by the summary column(s). The presentation of numerical values will adhere to the following guidelines:

- Raw measurements will be reported to the number of significant digits as captured electronically or on the CRFs.
- Rounding:
    - For derived parameters or raw results with more than 2 decimals:
        - Standard deviations will be reported to four decimal places
        - Means and medians will be reported to three decimal places
        - Minimums and maximums will be reported to two decimal places
    - For other parameters:
        - Standard deviations will be reported to two decimal places beyond the number of decimal places the original parameter is presented.
        - Means and medians will be reported to one decimal place beyond the number of decimal places the original parameter is presented.
        - Minimums and maximums will be reported to the same number of significant digits as the parameter.
- Calculated percentages will be reported with no decimals.

P-values will be presented to four decimal places. If a p-value is less than 0.0001 then it will be presented as "<0.0001".

11 References

Sansone SM, Schneider A, Bickel E, Berry-Kravis E, Prescott C, Hessl D. Improving IQ measurement in intellectual disabilities using true deviation from population norms. *J Neurodev Disord*. 2014;6(1):16.

ORCA Measure Scoring Manual. 2023. https://pattern.health/wp-content/uploads/2023/04/ORCA-Measure-Scoring-Manual_Final_2023_04_18.pdf