

**STATISTICAL ANALYSIS PLAN
For HGT-HIT-094 CSR**

Protocol Title: A Controlled, Randomized, Two-arm, Open-label, Assessor-blinded, Multicenter Study of Intrathecal Idursulfase-IT Administered in Conjunction with Elaprase® in Pediatric Patients with Hunter Syndrome and Early Cognitive Impairment

Protocol Number: HGT-HIT-094 (Amendment 5)

Protocol Date: 18 April 2016

Investigational Product and Device: idursulfase for intrathecal use (idursulfase-IT [HGT-2310]; SHP609), SOPH-A-PORT® Mini S, Implantable Access Port, Spinal, Mini Unattached, with Guidewire (SOPH-A-PORT® Mini S)

SAP Author: PPD [REDACTED], PPD [REDACTED], PPD [REDACTED]

SAP Version: Version 1.2

Release Date: 9November2017

Shire HGT Medical Monitor: PPD [REDACTED], DO
PPD [REDACTED]
300 Shire Way
Lexington, MA 02421, USA
Telephone: PPD [REDACTED]
Fax: PPD [REDACTED]

Confidentiality Statement

This document is the proprietary and confidential property of
Shire

HGT-HIT-094 STATISTICAL ANALYSIS PLAN APPROVAL SIGNATURES:

A Controlled, Randomized, Two-arm, Open-label, Assessor-blinded, Multicenter Study of Intrathecal Idursulfase-IT Administered in Conjunction with Elaprase® in Pediatric Patients with Hunter Syndrome and Early Cognitive Impairment

The signatures below indicate approval of the Statistical Analysis Plan for Protocol HGT-HIT-094 Amendment 5 dated 18 April 2016. Any changes or modifications to the Statistical Analysis Plan following approval, with the exception of minor editorial changes to table, figure or listing shells or clarification of shells for the programmers require an amendment with the corresponding approval signatures.

Author:

Yuna Wu, PhD

PPD [Redacted],
Global Clinical Operations
Shire
300 Shire Way
Lexington, MA 02421

PPD [Redacted]

Signature

PPD [Redacted]

Date
(DD MMM YYYY)

Reviewed and approved by:

Xiaoxi Li, PhD

PPD [Redacted],
Global Clinical Operations
Shire
300 Shire Way
Lexington, MA 02421

PPD [Redacted]

Signature

PPD [Redacted]

Date
(DD MMM YYYY)

David Alexanderian, DO

PPD [Redacted]
Shire
300 Shire Way
Lexington, MA 02421

PPD [Redacted]

Signature

PPD [Redacted]

Date
(DD MMM YYYY)

Table of Contents

1.	ABBREVIATIONS	6
2.	INTRODUCTION	8
2.1	Modification History	8
2.2	Background/Study Rationale.....	11
3.	PURPOSE OF THE STATISTICAL ANALYSIS PLAN	13
4.	SUMMARY OF CLINICAL TRIAL FEATURES	14
4.1	General Description.....	14
4.2	Determination of Sample Size/Randomization	19
5.	EFFICACY AND SAFETY ENDPOINTS	21
5.1	Schedule of Evaluations	21
5.2	Primary Efficacy Endpoint.....	21
5.3	Secondary Efficacy Endpoints	21
5.3.1	Key Secondary Efficacy Endpoint.....	21
5.3.2	Other Secondary Efficacy Endpoints.....	21
5.4	Exploratory Efficacy Endpoint.....	21
5.5	Pharmacokinetic and Pharmacodynamics Endpoints.....	21
5.6	Safety Assessments	21
5.7	SOPH-A-PORT® Mini S Device Assessments	21
5.8	Health Status Assessments	21
5.9	Cerebrospinal Fluid Assessments.....	21
6.	STATISTICAL ANALYSIS	23
6.1	General Methodology.....	23
6.2	Analysis Populations	23
6.3	Patient Disposition	24
6.4	Protocol Deviations.....	24
6.5	Demographic and Baseline Characteristics.....	24
6.6	Treatment Compliance and Extent of Exposure	25
6.7	Analysis of Efficacy	25
6.7.1	Primary Analysis.....	25
6.7.2	Key Secondary Analysis.....	26
6.7.3	Other Secondary Analyses.....	27
6.7.4	Subgroup Analyses	29
6.7.5	Exploratory Analyses.....	31
6.8	Analysis of Safety	33

6.8.1	Adverse Events	34
6.8.2	Clinical and CSF Laboratory Evaluations	35
6.8.3	12-Lead ECG Evaluations	35
6.8.4	Vital Signs.....	35
6.8.5	Physical and Neurological Examinations	36
6.8.6	Other Observations Related to Safety.....	36
6.8.7	Concomitant Medications	40
6.8.8	Concomitant Therapies	40
6.9	Analysis of Pharmacokinetic Data	40
6.9.1	Concentration Data	40
6.9.2	Handling BLQ Values.....	41
6.9.3	Pharmacokinetic Parameters.....	41
6.10	Drug-Drug and Drug-Disease Interactions	42
7.	CHANGES IN THE CONDUCT OF THE STUDY OR PLANNED ANALYSES.....	43
7.1	Changes in the Conduct of the Study	43
7.2	Changes from the Analyses Planned in the Protocol	43
8.	STATISTICAL/ANALYTIC ISSUES	44
8.1	Adjustment for Covariates	44
8.2	Handling of Dropouts or Missing Data.....	44
8.3	Interim Analyses and Data Monitoring.....	44
8.4	Multi-center Studies	44
8.5	Multiple Comparisons/Multiplicity.....	44
8.6	Examination of Subgroups and Interactions	46
8.7	Sensitivity Analyses	46
8.8	Windowing Visits.....	47
8.9	Data Listings	47
9.	APPENDICES	48
9.1	Appendix I - List of Statistical Outputs	48
9.2	Appendix II - Sample SAS code	48
9.2.1	Sample SAS code for the primary analysis.....	48
9.2.2	Sample SAS code for the key secondary analysis	48
9.2.3	Sample SAS code for the EQ-3D analysis.....	48
9.2.4	Sample SAS code for the correlation analysis using repeated measures.....	49
9.2.5	Sample SAS code for rate of change in GCA score analysis	49
9.2.6	Sample SAS code for the multiplicity analysis.....	50
9.2.7	Sample SAS code for sensitivity analysis using WOCF.....	50

9.2.8	Sample SAS code for sensitivity analysis using PMMs with control-based pattern imputation	50
9.2.9	Sample SAS code for the subgroup analysis with baseline age < 6 years, baseline age < 55 months and baseline age \geq 55 months	52
9.2.10	Sample SAS code for the subgroup analysis with baseline GCA classification factor	52
9.2.11	Sample SAS code for subgroup analysis testing treatment interaction	53
9.3	Appendix III - MMRM Simulation to Assess Analytical Methods	54
9.3.1	Introduction.....	54
9.3.2	Simulation Methods.....	54
9.3.3	Missing data.....	56
9.3.4	Results.....	56
9.3.5	Conclusions.....	59
9.3.6	Discussion.....	60
10.	REFERENCES	62

1. ABBREVIATIONS

ABC	Adaptive Behavior Composite
AE	Adverse event
ANCOVA	Analysis of Covariance
BBB	Blood brain barrier
BMI	Body mass index
BSID-III	Bayley Scales of Infant Development, 3rd edition
eCRF	electronic case report form
CI	Confidence interval
CSF	Cerebrospinal fluid
DAS-II	Differential Abilities Scale, 2nd edition
DQ	Developmental quotient
DMC	Data Monitoring Committee
EOS	end of study
EQ-5D	EuroQol-5D instrument for use as a measure of health outcome
FWER	family-wise type-I error rate
GAG	Glycosaminoglycan(s)
GCA	General Conceptual Ability
I2S	iduronate-2-sulfatase
ICP	intracranial pressure
IDDD	Intrathecal drug delivery device
ITT	Intent-to-Treat
LLOQ	lower limit of quantification
LP	Lumbar puncture
MedDRA	Medical Dictionary for Regulatory Activities
MI	Multiple Imputation
MMRM	Mixed Model Repeated Measures
MPS II	Mucopolysaccharidosis II (Hunter syndrome)
n/a	not applicable
PD	Pharmacodynamic(s)
PK	pharmacokinetic(s)
PMM	Pattern Mixture Model
PPD	Pharmaceutical Product Development, LLC
REML	restricted maximum likelihood estimation
SAE	serious adverse event
SAP	statistical analysis plan
SAS	Statistical Analysis System [®]
SD	standard deviation
SE	standard error
SEM	standard error of measurement
SNC	Special Nonverbal Composite
SOC	system organ class
SOPH-A-PORT Mini S	SOPH-A-PORT [®] Mini S, Implantable Access Port, Spinal, Mini Unattached, with Guidewire
SS	Standard score
vE	van Elteren test

WMW	Wilcoxon–Mann–Whitney test
WOCF	Worst observation carried forward
WRS	Wilcoxon Rank Sum

2. INTRODUCTION

2.1 Modification History

Unique Identifier for this SAP Version	Date of this SAP Version	Author	Major Changes from the Previous Version
1.1	03/04/2015	PPD PPD	<p>Removed the “Social-emotional” and “Adaptive behavior” from BSID-III because they were covered by VABS-II Added “Whole cortical volume (cm³)”</p> <p>Section 4.2: changed timing of BVA</p> <p>Section 8.7: per FDA recommendation, added sensitivity analyses to assess different covariance structures if the primary or key secondary models fail to converge.</p> <p>Section 6.7.1 & .2: Added sentence to check normality of the model. Section 8.7: Added additional sensitivity analyses. Section 6.7.4: Subgroup analyses by genotype and antibody status may be performed. Section 6.7.5.4: Analysis of both total GAG and HS GAG.</p> <p>Section 6.8.6.7.2: Minor corrections and clarifications to device performance analysis.</p> <p>IT device language updated per protocol amendment 3</p> <p>Section 6.9: Text added for planned PK analysis</p>
1.2	9November 2017	PPD	<p>Section 4.1: Re-ordered the section; added two exploratory efficacy endpoints; text added for additional efficacy endpoint for substudy only</p> <p>Section 4.1: Number of Subjects updated based on the BVA in Number of Subjects session and Randomization and Blinding</p>

			<p>session</p> <p>Section 4.1: Assessment schedule in Substudy added in the Treatment and Study Duration session</p> <p>Section 4.2: Text added and modified based on the BVA method</p> <p>Section 5: Text simplified and added section 5.4 for exploratory efficacy endpoints</p> <p>Section 6.2 Text modified for PK Population definition</p> <p>Section 6.7.1 Baseline age group (either <6 years or ≥ 6 years) and interaction terms added into the MMRM model; region removed from the MMRM model; sensitivity analysis was added</p> <p>Section 6.7.2 Baseline age group (either <6 years or ≥ 6 years) and interaction terms added into the MMRM model; region removed from the MMRM model</p> <p>Section 6.7.3 Text added for additional efficacy endpoint for substudy only; added MMRM model for 6 core subtests of DASII EY; added replaced Wilcoxon Rank Sum test with van Elteren test</p> <p>Section 6.7.4 Added subgroup analysis for subgroup with baseline age < 6 years, baseline age < 55 months and ≥ 55 months and baseline GCA classification factor</p> <p>Section 6.7.5.1 Replaced Wilcoxon Rank Sum test with van Elteren test</p> <p>Section 6.7.5.3 Removed content of BSID-II and content was moved to Section 6.7.3</p> <p>Section 6.7.5.4 Added rate of change in</p>
--	--	--	---

			<p>GCA score analysis</p> <p>Section 6.8.1 Text modified for TEAE definition</p> <p>Section 6.8.2 Box plot was replaced by individual patient plot</p> <p>Section 6.8.3 RR interval removed</p> <p>Section 6.8.4 Box plot was replaced by individual patient plot</p> <p>Section 6.8.6.1 Box plot was replaced by individual patient plot</p> <p>Section 6.8.6.2 Box plot was replaced by individual patient plot</p> <p>Section 6.8.6.5 Whole cortical volume was removed</p> <p>Section 6.8.6.7.1 Text modified for Device-related terminology and definition</p> <p>Section 6.8.6.7.2 Text modified for Device-related analysis</p> <p>Section 6.8.7 Text modified for concomitant medications definition</p> <p>Section 6.8.8 Text modified for concomitant therapies definition</p> <p>Section 6.9.1 Text modified for concentration data analysis</p> <p>Section 7.2 Added changes of analysis plan</p> <p>Section 8.2 Text removed for using median value imputation of MMRM model for handling of missing data</p> <p>Section 8.7 Baseline age group (either <6 years or ≥ 6 years) added to ANCOVA-</p>
--	--	--	---

			WOCF; replaced Wilcoxon Rank Sum test with van Elteren test Section 8.8 Text added for Windowing Visits Section 9.2 SAS code modified for interaction terms in the MMRM; SAS code modified for covariates in the MMRM and ANCOVA-WOCF model; SAS code added for slope analysis; SAS code added for sensitivity analysis using PMMs with control-based pattern imputation; SAS code added for subgroup analyses
--	--	--	--

2.2 Background/Study Rationale

Hunter syndrome is an extremely rare disease with an estimated incidence of 1 in approximately 162,000 live births worldwide (Meikle et al. 1999; Young and Harper, 1982). It is expected that approximately 67%-77% of these patients will present with the severe phenotype of the disease that includes progressive cognitive impairment (communicating hydrocephalus, increased intracranial pressure, seizures and hearing problems) and serious somatic disease.

The active ingredient of idursulfase-IT (for intrathecal use) drug product is idursulfase (recombinant human iduronate-2-sulfatase), which is the same active ingredient in commercially available Elaprase (for intravenous use). Elaprase is approved globally including the EU, the USA, Canada, and Japan as an intravenously (IV) administered enzyme replacement therapy (ERT) for patients with Hunter syndrome. Large proteins such as Elaprase are not expected to cross the blood brain barrier (BBB) in sufficient amounts to be therapeutically beneficial. Therefore, it is not possible to treat the progressive brain disease in severe Hunter syndrome with Elaprase, and direct administration of the active enzyme to the central nervous system is required. A formulation appropriate for intrathecal administration of idursulfase (idursulfase-IT [HGT-2310]), using an intrathecal drug delivery device (IDDD) the SOPH-A-PORT[®] Mini S (Implantable Access Port, Spinal, Mini Unattached, with Guidewire), has been developed to address this unmet medical need in an extremely rare population. The advantage of using an IDDD in a chronic disease such as Hunter syndrome is the potential to obviate the need for multiple lumbar punctures for drug delivery.

Nonclinical experience with IT administration of idursulfase-IT has demonstrated wide distribution of idursulfase to the CNS tissues. Idursulfase-IT has been shown to be well tolerated in several species and to be active in a murine disease model of idursulfase deficiency.

The non-interventional study HGT-HIT-050 was conducted to assist in identifying patients potentially eligible for enrollment in the Phase I/II study HGT-HIT-045.

In the completed Phase I/II clinical study HGT-HIT-045 and its ongoing extension study, HGT-HIT-046, idursulfase-IT has been generally well tolerated. Stabilization or improvement in cognitive and adaptive functions has been noted in some of the children enrolled in the trials. The available data support the Sponsor's hypothesis that a therapeutic benefit may be expected in MPS II children with cognitive impairment. The design of the Phase II/III study HGT-HIT-094, including the neurodevelopmental assessment tools and endpoints, and the selection of idursulfase-IT 10 mg dose, have been informed by the results to date from these Phase I/II studies.

The observational study HGT-HIT-090 was conducted to evaluate the neurodevelopmental status of pediatric MPS II patients over time and assist in identifying patients potentially eligible for enrollment in the Phase II/III study HGT-HIT-094. The duration of patient's participation in the study HGT-HIT-090 is up to 24 months. In the study HGT-HIT-094, the patients must meet the criteria that GCA standard score at screening is between 85 and 55. Patients with a GCA score at screening >85 who are ≥ 3 to ≤ 13 years of age may still be eligible to participate if there is demonstrated evidence of a decrease in GCA score of ≥ 10 points within a 12-month period in the observational study HGT-HIT-090.

3. PURPOSE OF THE STATISTICAL ANALYSIS PLAN

The purpose of this SAP is to document technical and detailed specifications for the final analysis of data collected from protocol HGT-HIT-094 Amendment 5 and to support completion of the clinical study report for submission to regulatory authorities.

4. SUMMARY OF CLINICAL TRIAL FEATURES

4.1 General Description

Study Design	<p>The pivotal study HGT-HIT-094 is a Phase II/III randomized, controlled, 2-arm, open-label, assessor-blinded, multicenter study to determine the effect on clinical parameters of neurodevelopmental status following monthly IT administration of idursulfase-IT 10 mg for 12 months in pediatric patients aged 3 to <18 years with Hunter syndrome and early cognitive impairment who have previously received and tolerated a minimum of 4 months of IV therapy with Elaprase (0.5 mg weekly). All patients continue to receive Elaprase therapy as standard of care throughout the study.</p> <p>The pivotal study design is “no IT treatment-controlled”. At least 48 patients are assigned randomly in a 2:1 ratio to IT treatment or no IT treatment arms.</p> <p>The separate substudy is open label with a single idursulfase-IT treatment arm. Patients will receive 12 monthly doses of idursulfase-IT. Patients in the substudy are <3 years old with Hunter syndrome and early cognitive impairment who have previously received and tolerated a minimum of 4 months of therapy with Elaprase. There is no target enrollment number for this substudy. The patients’ doses will be given based on reference brain weight as follows to account for brain weight development:</p> <ul style="list-style-type: none">• up to 8 month at dosing: 5 mg• >8 months to 30 months at dosing: 7.5 mg• >30 months to 3 years at dosing: 10 mg <p>Those patients in the substudy and those randomized to the IT treatment arm in the pivotal study will undergo surgical implantation of the SOPH-A-PORT[®] Mini S IDDD followed by a post-operative recovery period of at least 14 days prior to the first IT administration of idursulfase-IT. Treated patients in the pivotal study and the substudy will then receive 12 monthly IT injections of idursulfase-IT corresponding to a treatment and assessment interval of 13 months from randomization to the end-of-study (EOS) evaluations. Likewise, patients randomized to the no IT treatment arm will be assessed over 13 months after randomization.</p> <p>The SOPH-A-PORT[®] Mini S IDDD is used to obtain cerebrospinal fluid (CSF) samples and to deliver all IT injections of idursulfase-IT. If the IDDD appears to be non-functional, or if its use is precluded on a scheduled day of dosing, site personnel will refer to the IDDD Manual, which provides details on the investigation and management of any IDDD-related issues. If the IT space is not accessible via the IDDD, lumbar puncture may be utilized under defined circumstances for administration of idursulfase-IT or to obtain a CSF sample.</p>
--------------	---

Study Product	idursulfase for intrathecal use (idursulfase-IT [HGT-2310])
Intrathecal drug delivery device	SOPH-A-PORT [®] Mini S, Implantable Access Port, Spinal, Mini Unattached, with Guidewire (SOPH-A-PORT [®] Mini S)
Study Objectives	<p><i>Primary Objective</i></p> <ul style="list-style-type: none"> • To determine the effect of the treatment regimen in pediatric patients with Hunter syndrome and early cognitive impairment on the General Conceptual Ability (GCA) score as measured by the DAS-II, in conjunction with Elaprase therapy <p><i>Secondary Objectives</i></p> <p>The key secondary objective of this study is:</p> <ul style="list-style-type: none"> • To determine the effect of the treatment regimen in pediatric patients with Hunter syndrome and early cognitive impairment on the Adaptive Behavior Composite (ABC) score as measured by the VABS-II, in conjunction with Elaprase therapy <p>The secondary objectives of this study are:</p> <ul style="list-style-type: none"> • To determine the effect of the treatment regimen in pediatric patients with Hunter syndrome and early cognitive impairment, in conjunction with Elaprase therapy, on: <ul style="list-style-type: none"> ○ Cognitive function as measured by the cluster areas and subtests of the DAS-II ○ Adaptive behavior as measured by the domains and subdomains of the VABS-II <p><i>Safety Objectives</i></p> <ul style="list-style-type: none"> • To determine the effect of the treatment regimen on safety as assessed by adverse events, clinical laboratory testing, physical and neurological examination findings, vital signs, and electrocardiogram (ECG) results • To evaluate the anti-idursulfase antibody response in serum and cerebrospinal fluid (CSF) during the treatment regimen <p><i>SOPH-A-PORT[®] Mini S Device Objectives</i></p> <ul style="list-style-type: none"> • To determine the safety and performance of the SOPH-A-PORT[®] Mini S device <p><i>Pharmacokinetic and Pharmacodynamic Objectives</i></p> <ul style="list-style-type: none"> • To evaluate the concentration of idursulfase and determine pharmacokinetic (PK) parameters in serum after IT administration • To evaluate the concentration of idursulfase in CSF prior to each monthly administration of idursulfase-IT • To determine the effect of the treatment regimen on the concentration of

	<p>glycosaminoglycan (GAG) in CSF</p> <p><i>Health Status Objective</i></p> <ul style="list-style-type: none"> To evaluate health status as measured by the EQ-5D instrument <p><i>Substudy Objective</i></p> <ul style="list-style-type: none"> To examine the effect of the treatment regimen on safety and efficacy measures in pediatric patients with Hunter syndrome and early cognitive impairment who are below 3 years of age
<p>Study Endpoints</p>	<p><i>Primary efficacy endpoint</i></p> <ul style="list-style-type: none"> Change from baseline in the GCA standard score after 12 months of treatment at Visit Week 52, as obtained by DAS-II testing <p><i>Key secondary efficacy endpoint</i></p> <ul style="list-style-type: none"> Change from baseline in the ABC standard score after 12 months of treatment at Visit Week 52, as obtained by VABS-II testing <p><i>Secondary efficacy endpoints</i></p> <ul style="list-style-type: none"> Change from baseline to Visit Weeks 16, 28, and 40 in the GCA standard score as obtained by DAS-II testing Change from baseline to Visit Weeks 16, 28, and 40 in the ABC standard score as obtained by VABS-II testing Change from baseline to Visit Weeks 16, 28, 40, and 52 in standard scores in cluster areas of the DAS-II: Verbal, Nonverbal, Spatial, and Special Nonverbal Composite (SNC) Change from baseline to Visit Weeks 16, 28, 40, and 52 in the standard scores of the VABS-II domains: Communication, Daily Living Skills, Socialization, and Motor Skills Change from baseline to Visit Weeks 16, 28, 40, and 52 in the age equivalents, developmental quotients (DQ), and T scores for the subtests of the DAS-II (Verbal Comprehension, Picture Similarities, Naming Vocabulary, Pattern Construction, Matrices, and Copying for the DAS-II/Early Years, and Recall of Designs, Word Definitions, Pattern Construction, Matrices, Verbal Similarities, and Sequential and Quantitative Reasoning for the DAS-II/School Years) Change from baseline to Visit Weeks 16, 28, 40, and 52 in the age equivalents, DQ and v-scale scores of the VABS-II subdomains: Communication (Receptive, Expressive, Written), Daily Living Skills (Personal, Domestic, Community), Socialization (Interpersonal Relationships, Play and Leisure Time, Coping Skills), Motor Skills (Gross, Fine) Change from baseline to Visit Weeks 16, 28, 40, and 52 in v-scale scores and the observed maladaptive levels of the VABS-II Maladaptive

	<p>Behavior Index and its subscales (Internalizing, Externalizing)</p> <p><i>Exploratory efficacy endpoint</i></p> <ul style="list-style-type: none"> • Slope estimate from the linear regression of GCA scores over time for each subject <p><i>Additional efficacy endpoint for substudy only</i></p> <ul style="list-style-type: none"> • Observed value in raw score, age equivalents versus chronological age, as well as DQ, composite score and percentile for the three scales of the BSID-III (Motor [Fine, Gross], Language [Receptive, Expressive], Cognitive) for substudy only <p><i>Pharmacokinetic and Pharmacodynamic endpoints</i></p> <ul style="list-style-type: none"> • Serum concentration of idursulfase and serum PK parameters after IT administration • CSF concentration of idursulfase prior to each monthly IT administration • Change from baseline in the concentration of GAG in CSF <p><i>Safety Assessments</i></p> <p>Safety will be assessed during the study by collection of adverse events (by type, severity, and relationship to treatment [idursulfase-IT, the IDDD, device surgical procedure, or IT administration process] and IV Elaprase infusion), changes in clinical laboratory testing (serum chemistry, hematology, urinalysis), physical and neurological examination, vital signs, 12-lead ECG, CSF laboratory parameters (chemistries, cell counts), anti-idursulfase antibodies in CSF and serum, and determination of antibodies having enzyme neutralizing activity.</p> <p><i>SOPH-A-PORT[®] Mini S Device Assessments</i></p> <p>The SOPH-A-PORT Mini S device will be evaluated using assessments of device implantation, device function, device longevity and adverse events associated with the implant surgery or device. Those data will be collected on the patient's electronic case report form (eCRF) from the time of initial implantation and continue throughout the study as long as the SOPH-A-PORT[®] Mini S remains implanted.</p> <p><i>Health Status Assessment</i></p> <p>Health status dimensions as obtained by the EQ-5D questionnaire: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression.</p>
Number of Subjects	<p>Pivotal study</p> <ul style="list-style-type: none"> • 48 male patients (32 IT treated and 16 No IT treatment) are planned.

	<p>Substudy</p> <ul style="list-style-type: none"> • There is no pre-specified number of patients planned. The enrollment will conclude when enrollment of patients in the pivotal study closes.
<p>Treatment and Study Duration</p>	<p>Patients in the IT treatment arm of the pivotal study will be assessed according to the following schedule:</p> <ul style="list-style-type: none"> • Screening (Visit Weeks -4 to -1 [Day -28 to Day -1]) • Randomization (Visit Week 0 [Day 0]) • Pre-surgery, Surgery, Follow-up, and Post-operative Recovery (Visit Week 2 [+7 days]) • Treatment and Assessments(Visit Week 4 through Week 48 [±7 days]) • End of study (EOS, Visit Week 52 [±7 days]) • Follow-up (telephone contact) 7 (±2) days from the Visit Week 52 (or EOS) <p>Patients in the no IT treatment arm of the pivotal study will be assessed according to the following schedule:</p> <ul style="list-style-type: none"> • Screening (Visit Weeks -4 to -1 [Day -28 to Day -1]) • Randomization (Visit Week 0 [Day 0]) • Telephone contact (Week 2 [+7 days]) • Assessments (Visit Week 4 through Week 48 [±7 days]) • End of study (EOS, Visit Week 52 [±7 days]) • Follow-up (telephone contact) 7 (±2) days from the Visit Week 52 (or EOS) <p>The separate substudy is open-label and single arm. Patients who meet all entry criteria for participation in the substudy will be considered enrolled on Day 0. Thereafter, patients who are enrolled in the separate substudy will follow a similar schedule of treatment and assessments as idursulfase-IT-treated patients in the pivotal study.</p> <p>Patients in the substudy will be assessed according to the following schedule:</p> <ul style="list-style-type: none"> • Screening (Visit Weeks -4 to -1 [Day -28 to Day -1]) • Enrollment (Week 0 [Day 0]) • Pre-surgery, Surgery, Follow-up, and Post-operative Recovery (Visit Week 2 [+7 days]) • Treatment and Assessments(Visit Week 4 through Week 48 [±7 days]) • End of study (EOS, Visit Week 52 [±7 days]) • Follow-up (telephone contact) 7 (±2) days from the Visit Week 52 (or EOS) <p>The planned overall duration of each patient’s participation in the study is approximately 14 to 15 months from Screening to the last scheduled contact.</p>
<p>Randomization and</p>	<p>A centralized randomization procedure is utilized to randomize patients in a 2:1</p>

Blinding	<p>allocation ratio to study arms (32 IT treated and 16 No IT treatment patients). The randomization is stratified by baseline GCA score (≤ 70 and > 70).</p> <p>HGT-HIT-094 is an open label study. Single and double blinding of patients, their families, and the Principal Investigator is not possible due to the absence of a sham device, sham injections, or placebo. The assessors responsible for the DAS-II and VABS-II evaluations will not be informed of patients' randomization assignments. The families will be instructed not to share this information with the assessors.</p> <p>Different assessors will be responsible for administration of the DAS-II and VABS-II. All assessments will be administered by qualified study personnel. The DAS-II will be administered by a trained practitioner. Shire will have no access to unblinded DAS-II and VABS-II data during the study.</p> <p>The separate substudy is open label and non-randomized, with a single idursulfase-IT treated arm.</p>
----------	--

4.2 Determination of Sample Size/Randomization

Based on study HGT-HIT-050/045 Phase I/II data, the observed decline in mean General Conceptual Ability (GCA) standard score (SS) in No IT treatment patients was estimated to be 13.7 points with a standard deviation of 9.4 after 12 months. For the sample size calculation, we conservatively assumed an approximately 1 point per month rate of decline in the No IT treatment arm, so that the theoretical decline would be 13 points from baseline (Screening Visit) to Visit Week 52 (~ 13 months). As up to 2 months may elapse between the screening assessment and the start of idursulfase-IT treatment, a mean decline of 2 points at end of study would be expected if idursulfase was effective in stabilizing the decline. Therefore, in such a case, the mean projected treatment difference at Visit Week 52 would be 11 points, which is considered a clinically meaningful treatment difference.

Using a 2:1 allocation ratio, a sample size of 48 randomized patients (about 32 IT treated patients and 16 no IT treatment patients) will yield 80% power to detect a clinically meaningful mean treatment difference of 11 points in the primary endpoint, GCA change from baseline to Visit Week 52. This calculation further assumes a common standard deviation for the change from baseline of 11.6 points, a type-I error rate of 0.05 for a two-sided two-sample t-test, with approximately up to 10% missing an assessment at Visit Week 52. Given this sample size, and conditional on rejection of the null hypothesis for the primary endpoint, the power would be 80% to detect a clinically meaningful mean difference of 10 points in the key secondary endpoint, change from baseline in ABC standard score at Visit Week 52, assuming a standard deviation for the change from baseline of 10.5 points for a two-sided two-sample t-test with a significance level of 5% and 10% missing an assessment at Visit Week 52.

The variance estimate for this sample size was determined based on a blinded variability assessment (BVA) performed as specified in the prior protocol version (Amendment 3); refer to

[Appendix 5](#) of the protocol Amendment 5 for details. Given that uncertainty around this variability estimate remained, a second BVA using the same approach as specified in [Appendix 5](#) of the protocol Amendment 5 was performed to confirm the estimate after approximately 27 patients completed the Week 52 primary endpoint assessment. Based on this second blinded assessment, the sample size was set to 48 patients while maintaining the 2:1 allocation ratio. As there was no treatment comparison involved, no inflation of the type I error from this procedure was expected, and no adjustment to the significance levels for the final analysis was necessary.

Patients will be randomized (open-label, assessor-blinded) in a 2:1 allocation ratio between the idursulfase-IT treatment regimen and no IT treatment groups. Although this is a multi-center study, the randomization will not be stratified by center. The baseline GCA score (at Screening Visit) is expected to be a key prognostic factor. Therefore, the randomization will be stratified according to baseline GCA score: ≤ 70 versus > 70 . The randomization schedule will be generated and administered centrally by Pharmaceutical Product Development, LLC (PPD), independent of the sponsor's project team. There is no replacement of patients who do not complete the study.

In a separate substudy, additional patients with cognitive impairment may be enrolled who are below the age of 3 years at the time of informed consent. All patients meeting eligibility criteria for the substudy will be enrolled until enrollment for the pivotal study is completed. There is no target sample size. Such patients are ineligible for the pivotal trial and will not be randomized, but will receive idursulfase-IT treatment monthly.

5. EFFICACY AND SAFETY ENDPOINTS

5.1 Schedule of Evaluations

The detailed schedule of evaluations can be found in the HGT-HIT-094 protocol (Amendment 5).

5.2 Primary Efficacy Endpoint

The primary efficacy endpoint is described above in Section 4.1.

5.3 Secondary Efficacy Endpoints

5.3.1 Key Secondary Efficacy Endpoint

The key secondary efficacy endpoint is described above in Section 4.1.

5.3.2 Other Secondary Efficacy Endpoints

The other secondary efficacy endpoints are described above in Section 4.1.

5.4 Exploratory Efficacy Endpoint

The exploratory endpoint is described above in Sections 4.1.

5.5 Pharmacokinetic and Pharmacodynamics Endpoints

The PK and PD endpoints are described above in Section 4.1.

5.6 Safety Assessments

The safety assessments are described above in Section 4.1.

5.7 SOPH-A-PORT® Mini S Device Assessments

The device assessments are described above in Section 4.1.

5.8 Health Status Assessments

The health status assessments are described above in Section 4.1.

5.9 Cerebrospinal Fluid Assessments

Cerebrospinal fluid will be obtained from patients in the treated arm of the pivotal study and in the substudy at Screening Visit (by lumbar puncture and under general anesthesia), during surgical implantation of the IDDD, prior to each intrathecal injection of investigational product, and at the Week 52/EOS visit. Patients in the no treatment arm of the pivotal study will undergo

lumbar puncture (under general anesthesia) to obtain CSF samples at Screening and the Week 52/EOS visits.

6. STATISTICAL ANALYSIS

6.1 General Methodology

All statistical analyses will be performed by the Biostatistics and Statistical Programming department at Shire unless otherwise specified, using SAS[®] software version 9.3 or higher (SAS Institute, Cary, N.C., USA).

Continuous variables will be summarized using descriptive statistics (N, mean, standard deviation, minimum, median, and maximum). Categorical variables will be summarized using the number and percentage of patients in each category. Data will be summarized with respect to patient disposition, demographic and baseline characteristics and concomitant medication use. The efficacy endpoints, safety assessments and other outcome results for each treatment group will be summarized descriptively unless otherwise indicated. In addition, statistical model estimates of least squares means, treatment differences, p-values and 95% confidence intervals for least squares mean treatment differences will also be provided where relevant for efficacy endpoints. The fit of linear models will be assessed using residual plots and/or other diagnostic plots as appropriate. All the statistical tests will be 2-sided and performed at the 0.05 level of significance unless stated otherwise. Baseline is defined as the value at Screening Visit for all parameters, unless specified otherwise. The planned analyses described below in this SAP pertain to the pivotal study data.

The separate substudy is an open-label, single arm study. Data from patients participating in this separate substudy will be analyzed separately. Since the number of patients enrolled in the substudy is expected to be low, their safety and efficacy data will not be summarized, but the corresponding data as described for the pivotal study will be listed. The listing format would be the same as that for the pivotal study where relevant.

6.2 Analysis Populations

For the pivotal study, all efficacy and pharmacodynamic data analyses will be performed according to treatment assigned using the Intent-to-Treat (ITT) Population, which is defined as all randomized patients.

All safety data analyses will be performed according to treatment received using the Safety Population, which is defined as all randomized patients with any post-randomization safety assessments. IDDD and procedure related analyses will be conducted in the set of patients in the Safety Population who had the device implant procedure performed.

For the substudy, analyses will be performed on the Substudy Population, defined as all patients enrolled and treated with study drug in the substudy.

All pharmacokinetic data analyses will be performed using the Pharmacokinetic (PK) Population. The PK Population is defined as all patients who receive study drug and participate

in the scheduled pharmacokinetic studies, and for whom at least 1 post-dose PK blood sample was collected.

6.3 Patient Disposition

The number of patients screened (i.e., signed informed consent) will be presented. The number and percentage of patients randomized, patient population (ITT, Safety Populations, and PK Population), and treatment status (completed, discontinued / withdrew) will be summarized by treatment group and overall. Reasons for discontinuation/withdrawal will be presented.

6.4 Protocol Deviations

An incident involving noncompliance with the protocol, but one which typically does not have significant effects on the patient's rights, safety, or welfare, or the integrity of the resultant data will be considered a protocol deviation. An incident involving noncompliance with the protocol which may affect the patient's rights, safety, or welfare, or the integrity of the resultant data will be known as a protocol violation. In particular, any serious deviation that affects the collection of data for the primary endpoint will be considered a protocol violation. In general, protocol exemptions will not be granted.

Examples of potential protocol violations may include violation of important admission (inclusion/exclusion) criteria, occurrence of a treatment dispensing error, treatment noncompliance or substantial use of a prohibited medication during the study. Reported protocol deviations and patient data will be examined prior to database lock to determine if conditions set forth in the study protocol have been violated and a more comprehensive list will be constructed at that time. The complete list of major protocol deviations will not be summarized; however, a list of major protocol deviation will be presented.

6.5 Demographic and Baseline Characteristics

Demographic and baseline characteristics (e.g., age [years], race, ethnicity, weight [kg], height [cm], BMI [kg/m²], genotype, country, age at onset of Hunter syndrome symptoms, age of Hunter syndrome diagnosis, family history of Hunter syndrome) will be summarized by treatment group and overall. The genotype will be summarized into the following categories:

- Complex rearrangement
- Frameshift
- Missense
- Nonsense
- Splice site mutation

$BMI = \text{weight (kg)} / (\text{height (m)})^2$. $\text{Age} = (\text{date of event} - \text{birth date} + 1) / 365.25$.

Date of event is defined as randomization date for the pivotal study patients, and date of eligibility (enrollment date) for the substudy patients. If only the event day is missing, the day

will be assigned as 15. If both the day and month are missing, they will be assigned as July 1st. One decimal will be shown for BMI and age.

Medical history will be coded by the Medical Dictionary for Regulatory Activities (MedDRA) version 16.1 or higher and surgical history will be coded by the Shire Medical Monitor. This data will be summarized by treatment group.

6.6 Treatment Compliance and Extent of Exposure

The number of IT injections received overall and by lumbar puncture, average dose, the duration of idursulfase-IT treatment and IT administration duration will be summarized descriptively in the idursulfase-IT arm.

The duration of idursulfase-IT treatment, summarized in months, is defined as the time from first to last IT administration during the study.

The duration for each idursulfase-IT administration (in minutes) is calculated by subtracting the administration start time from the administration end time. The actual average duration of idursulfase-IT administration (in minutes) will be calculated as an average value of the idursulfase-IT administrations across all available administrations for the same patient.

Treatment compliance will be summarized in terms of the percent of scheduled doses received in the idursulfase-IT arm. The percent of scheduled doses received for each patient is defined as: $[(\text{Number of Complete IT injections Received}) \div (\text{Expected Number of IT injections at EOS})] * 100$.

6.7 Analysis of Efficacy

All efficacy analyses described below refer to the pivotal study; all efficacy data from patients treated in the separate substudy will be listed, including measurements from the BSID-III, VABS-II, PD data and DAS-II if available. All change from baseline endpoints are calculated as the value of the corresponding visit minus the value at baseline.

6.7.1 Primary Analysis

For the pivotal study, the primary efficacy endpoint is the change from study baseline (Screening Visit) to Visit Week 52 in the General Conceptual Ability (GCA) score as measured by the DAS-II. The primary analysis will compare the treatment groups using a linear mixed model repeated measures (MMRM) analysis (Malinckrodt et al. 2008; Siddiqui et al. 2009). The repeated measures are the change from baseline GCA score obtained at the scheduled Visit Weeks 16, 28, 40, and 52, respectively. The model will include fixed categorical effects for treatment, visit week, treatment by visit week interaction, baseline GCA classification factor (either ≤ 70 or > 70), baseline age group (either < 6 years or ≥ 6 years), treatment by baseline GCA classification factor interaction, treatment by baseline age group interaction, interaction between baseline GCA classification factor and baseline age group, genotype, and the baseline GCA score as a continuous covariate. Visit week is defined as nominal visits, unless specified otherwise.

SAS Proc Mixed with restricted maximum likelihood estimation (REML) and an unstructured within-patient covariance structure will be used. If this model fails to converge, a first order autoregressive (AR(1)) covariance structure will be used for the primary analysis. Due to small sample size of ITT Population with baseline age ≥ 6 years, MMRM model may fail to converge when low count of subjects with baseline age ≥ 6 years exists in either treatment group. If this model fails to converge, the treatment by baseline age group interaction and /or interaction between baseline GCA classification factor and baseline age group will be removed from the model. The Kenward-Roger approximation will be used to estimate denominator degrees of freedom for tests of fixed effects. The assumptions of the model, including normality, will be evaluated using residual and other diagnostic plots of model fit.

From this model, least squares means, standard errors, treatment differences in least squares means, and 95% confidence intervals will be estimated for each time point. Primary inference will be based on the treatment comparison of least squares means at Visit Week 52 from this model, and a p-value will be presented for this time point only. The null hypothesis is that the mean difference in the primary endpoint between the two treatment groups is zero, versus the alternative hypothesis that this difference is not zero. The hypotheses can be expressed as follows:

$$H_0: \mu_{IT} - \mu_{control} = 0 \text{ versus } H_1: \mu_{IT} - \mu_{control} \neq 0$$

Where μ_{IT} refers to the mean change from baseline to Visit Week 52 in GCA score in the idursulfase-IT treated group and $\mu_{control}$ refers to the mean change from baseline to Visit Week 52 in GCA score in the No IT treatment group. The test will be performed using the final, MMRM model-based t-test with a two-sided significance level of 5%. Estimated least squares means for change from baseline and the observed absolute values (\pm SE) by treatment group will be plotted over time.

The sample SAS code is presented in [appendix 9.2.1](#).

As a sensitivity analysis for the primary analysis using ITT population, MMRM model in the same manner as described above will be conducted for a subset of ITT population, which includes all ITT subjects excluding subjects enrolled from the observational study HGT-HIT-090 who experienced a decrease in GCA score of ≥ 10 points over 12 months in study HGT-HIT-090 prior to being enrolled in study HGT-HIT-094 and a GCA score at HGT-HIT-094 Screening >85 points.

6.7.2 Key Secondary Analysis

The key secondary efficacy endpoint is the change from study baseline (Screening Visit) to Visit Week 52 in the Adaptive Behavior Composite (ABC) score as measured by the VABS-II. The key secondary analysis will compare the treatment groups using a linear mixed model repeated measures (MMRM) analysis. The repeated measures are the change from baseline ABC scores obtained at the scheduled Visit Weeks 16, 28, 40, and 52, respectively. The model will include fixed categorical effects for treatment, visit week, treatment by visit week interaction, baseline GCA classification factor (either ≤ 70 or >70), baseline age group (either <6 years or ≥ 6 years),

treatment by baseline GCA classification factor interaction, treatment by baseline age group interaction, interaction between baseline GCA classification factor and baseline age group, genotype, and the baseline ABC score as a continuous covariate.

SAS Proc Mixed with restricted maximum likelihood estimation (REML) and an unstructured within-patient covariance structure will be used. If this model fails to converge, a first order autoregressive (AR(1)) covariance structure will be used instead. The Kenward-Roger approximation will be used to estimate denominator degrees of freedom for tests of fixed effects. The assumptions of the model, including normality, will be evaluated using residual and other diagnostic plots of model fit.

From this model, least squares means, standard errors, treatment differences in least squares means, and 95% confidence intervals will be estimated for each time point. The significance test will be based on the treatment comparison of least squares means at Visit Week 52 from this model, and a p-value will be presented for this time point only. Estimated least squares means for change from baseline and the observed absolute values (\pm SE) by treatment group will be plotted over time.

The sample SAS code is presented in [appendix 9.2.2](#).

6.7.3 Other Secondary Analyses

The following secondary efficacy endpoints will be analyzed using a MMRM analysis in the same manner as described above for the key secondary endpoint with the continuous covariate corresponding to the baseline score for each measure:

- The change from baseline to Visit Weeks 16, 28, 40, and 52 in standard scores in cluster areas of the DAS-II: Verbal, Nonverbal, Spatial, and Special Nonverbal Composite (SNC)
- The change from baseline to Visit Weeks 16, 28, 40, and 52 in the standard scores of the VABS-II: Communication, Daily Living Skills, Socialization, and Motor Skills
- The change from baseline to Visit Weeks 16, 28, 40, and 52 in the T scores for the six core subtests of the DAS-II/Early Years: Verbal Comprehension, Picture Similarities, Naming Vocabulary, Pattern Construction, Matrices and Copying

Due to small sample size of ITT subjects who have been evaluated by the DAS-II Test for school age, the six core subtests for the DAS-II/School Age, including Recall of Designs, Word Definitions, Pattern Construction, Matrices, Verbal Similarities and Sequential and Quantitative Reasoning, cannot be analyzed using the MMRM model.

The estimated least squares means and 95% confidence intervals on the treatment comparison at Visit Weeks 16, 28, 40 and 52 from MMRM model will be presented in Forest plots for GCA score, standard scores for SNC and each of the three clusters (Verbal, Nonverbal, Spatial), and T scores for each of the six above-mentioned core subtests of the DAS-II/Early Years.

All other secondary efficacy endpoints listed above will be summarized descriptively by treatment group over time. When there is $>$ or \geq or $<$ or \leq recorded in the data, the inequality sign will be ignored for the age equivalents summary and the DQ calculation. DQ is defined as Age equivalent provided by the test $\times 100$ /Chronological age.

Mean observed values by treatment group will be plotted over time. This includes the following endpoints:

- The change from baseline to Visit Weeks 16, 28, 40, and 52 in the T scores, age equivalents and DQ for the core subtests of the DAS-II: Verbal Comprehension, Picture Similarities, Naming Vocabulary, Pattern Construction, Matrices and Copying for the DAS-II/Early Years; and Recall of Designs, Word Definitions, Pattern Construction, Matrices, Verbal Similarities and Sequential and Quantitative Reasoning for the DAS-II/School Years
- The change from baseline to Visit Weeks 16, 28, 40, and 52 in the v-scale scores, age equivalents and DQ for the sub-domains of the VABS-II: Communication (Receptive, Expressive, Written), Daily Living Skills (Personal, Domestic, Community), Socialization (Interpersonal Relationships, Play and Leisure Time, Coping Skills), Motor Skills (Gross, Fine)
- Change from baseline to Visit Weeks 16, 28, 40, and 52 in v-scale scores and the observed maladaptive levels of the VABS-II Maladaptive Behavior Index and its sub-scales (Internalizing, Externalizing).

By-patient listings will be presented for the following parameters:

- Standard scores and percentiles in patients evaluated by the DAS-II test (either version) for each cluster
 - Verbal
 - Nonverbal
 - Spatial
 - GCA
 - SNC
- T scores, percentiles of T scores, age equivalents, chronological age and DQ for patients evaluated by the DAS-II test for Early Years for each core subtest
 - Verbal Comprehension (Vcom)
 - Picture Similarities (Psim)
 - Naming Vocabulary (Nvoc)
 - Pattern Construction (Pcon)
 - Matrices (Mat)
 - Copying (Copy)
- T scores, percentiles of T scores, age equivalents, chronological age and DQ for patients evaluated by the DAS-II test for School Age for each core subtest
 - Recall of Designs (RDes)

- Word Definitions (WDef)
- Pattern Construction (Pcon)
- Matrices (Mat)
- Verbal Similarities (VSim)
- Sequential & Quantitative Reasoning (SQR)

- Standard scores and percentile in patients evaluated by the VABS-II test for ABC and each domain
 - Communication
 - Daily Living Skills
 - Socialization
 - Motor Skills

- v-scale scores, age equivalents, chronological age and DQ for patients evaluated by the VABS-II test for each sub-domain
 - Communication (Receptive, Expressive, Written),
 - Daily Living Skills (Personal, Domestic, Community),
 - Socialization (Interpersonal Relationships, Play and Leisure Time, Coping Skills),
 - Motor Skills (Gross, Fine)

- v-scale scores and Maladaptive levels of Maladaptive Behavior Index and its sub-scales (Internalizing, Externalizing) in VABS-II

A listing for the substudy population will be presented by patient and visit for the following parameters:

- Raw score, age equivalents versus chronological age, as well as DQ, composite score and percentile for each patient evaluated by the BSID-III test and each domain
 - Motor (Fine, Gross)
 - Language (Receptive, Expressive)
 - Cognitive

6.7.4 Subgroup Analyses

Subgroup analyses will be performed for the following subgroups:

- Baseline GCA classification groups. For this analysis, patients with a baseline GCA score of > 70 will be classified as having “Moderate” cognitive impairment, while patients with a baseline GCA score ≤ 70 will be classified as having “Severe” cognitive impairment. The randomization is stratified by this classification variable to ensure treatment group balance within these subgroups
- Baseline age group (either < 6 years or ≥ 6 years)
- Exploratory analysis for baseline age group (either < 55 months or ≥ 55 months).

Subgroup analyses are planned for the change from baseline in GCA and ABC standard scores, respectively. An MMRM model will be used to test for treatment by subgroup interactions. Interactions with a significance level of less than 10% will be considered potentially important and flagged for further clinical assessment. In general, the model will include fixed categorical effects for treatment, visit week, treatment by visit week interaction, subgroup, treatment by subgroup interaction, subgroup by visit interaction and the 3-way interaction between treatment, visit week and subgroup, and genotype. The p-values of interaction terms will be presented, as well as the least squares means and its 95% confidence intervals by treatment and visit for each subgroup classification factor. Descriptive statistics of the observed and change from baseline GCA and ABC standard scores will also be presented by treatment and visit week within each subgroup.

Additionally, the subgroup analysis will be performed for the age subgroup with baseline age < 6 years which may be the patient population with the most potential to benefit from the idursulfase-IT treatment. The change from baseline GCA score obtained at Visit Week 52 will be compared between the treatment groups using MMRM analysis. The repeated measures are the change from baseline GCA score obtained at the scheduled Visit Weeks 16, 28, 40, and 52, respectively. The model will include fixed categorical effects for treatment, visit week, treatment by visit week interaction, baseline GCA classification factor (either ≤ 70 or > 70), treatment by baseline GCA classification factor interaction, genotype and the baseline GCA score as a continuous covariate. Due to small sample size of ITT subjects with baseline age ≥ 6 years, MMRM model will not be applied for this age subgroup.

The sample SAS code is presented in [appendix 9.2.9](#).

As exploratory analysis, the subgroup analysis will also be performed for the age subgroup with baseline age < 55 months and ≥ 55 months using an MMRM analysis in the same manner as described above for age subgroup with baseline age < 6 years.

The subgroup analyses will also be performed for the baseline GCA subgroup with baseline GCA score ≤ 70 and baseline GCA score > 70 . The change from baseline GCA score obtained at Visit Week 52 will be compared between the treatment groups using MMRM analysis. The repeated measures are the change from baseline GCA score obtained at the scheduled Visit Weeks 16, 28, 40, and 52, respectively. The model will include fixed categorical effects for treatment, visit week, treatment by visit week interaction, baseline age group (either <6 years or ≥ 6 years), treatment by baseline age group interaction, genotype and the baseline GCA score as a continuous covariate.

The sample SAS code is presented in [appendix 9.2.10](#).

The estimated least squares means and 95% confidence intervals on the treatment comparison at Visit Weeks 16, 28, 40, and 52 from MMRM model will be presented in Forest plots, including subgroup of baseline age < 6 years, baseline age < 55 months and baseline age ≥ 55 months, baseline GCA score ≤ 70 and baseline GCA score > 70 .

As exploratory analyses, the primary and key secondary MMRM models will be refit including the continuous baseline covariate by treatment interaction term and the 3-way interaction among treatment, visit week and the continuous baseline covariate in the model. The p value for the interaction terms and the estimated least squares means with their standard errors and 95% confidence intervals by treatment and visit at quartiles of GCA at baseline (25th, 50th and 75th percentiles) will be presented.

Due to small sample size of ITT Population with baseline age ≥ 6 years, MMRM model may fail to converge when low count of subjects with baseline age ≥ 6 years exists in either treatment group. If this model fails to converge, the treatment by baseline age group interaction will be removed from the model.

The sample SAS code is presented in [appendix 9.2.11](#).

6.7.5 Exploratory Analyses

6.7.5.1 Health Status Outcome

The health status of patients will be assessed using the EQ-5D questionnaire. The EQ-5D measures 5 dimensions of health status: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. For each dimension, there are 5 levels of response:

- No problems
- Slight problems
- Moderate problems
- Severe problems
- Unable to do / Extreme problems

The number and percent of patients with each response will be presented by dimension at Screening visit and Visit Week 52. The visual analogue scale (VAS) records the patient's parent/caregiver-rated health on a 0 (worst health) to 100 (best health) scale. The VAS score, as well as the change from baseline (Screening visit to Visit Week 52) will be summarized.

A van Elteren (vE) test ([van Elteren 1960](#)) will be used to compare treatment groups on the change from baseline to Visit Week 52 in VAS score.

The Wilcoxon–Mann–Whitney (WMW) test ([Mann and Whitney 1947](#); [Wilcoxon 1945](#)) is widely used nonparametric approach to compare two treatments. In the presence of stratum factors such as baseline age group or baseline cognitive impairment status, the van Elteren (vE) test ([van Elteren 1960](#)), a stratified WMW test, can be used to adjust for the stratification factors.

The treatment effect on change from baseline to Visit Week 52 in VAS score will be tested by the van Elteren test, which is a direct extension of the 2-sample WMW test, and a particular form of the Cochran-Mantel-Haenszel (CMH) test. Baseline age group (either <6 years or ≥ 6 years) and baseline GCA score (either ≤ 70 or >70) will be used as stratification factors in the van Elteren test. It is shown by Qu et al. ([2008](#)) that vE test preserves the type I error rate regardless

of the existence of the stratum effects, and is better than the WMW test when the stratum effects are large.

The null hypothesis being tested is:

H_0 : Two treatment groups produce the same distribution of outcomes

Using one-sided test at the 0.025 level, rejection of H_0 would indicate that the IT treatment group produces more desirable outcomes than no IT treatment group.

The sample SAS code is presented in [appendix 9.2.3](#).

Pharmacoeconomic analyses may be performed by the sponsor staff in the Outcomes Research and Epidemiology group and reported separately in a pharmacoeconomic report to be appended to the Clinical Study Report. Accordingly, any planned pharmacoeconomic analyses related to this data may be described elsewhere.

6.7.5.2 Correlation between Different Neurodevelopmental Assessment Domains

The following correlation coefficients for observed values and change from baseline will be estimated from a linear mixed effects model that takes into account the repeated measures on each patient using the method described by Hamlett et al. (2003) SAS Proc Mixed with restricted maximum likelihood estimation (REML) and an unstructured between and within-patient covariance structure for the random effects will be used for this model.

- GCA standard score from DAS-II and
 - Verbal standard score from DAS-II
 - Nonverbal standard score from DAS-II
 - Spatial standard score from DAS-II
 - SNC standard score from DAS-II
 - ABC standard score from VABS-II
 - Communication standard score from VABS-II
 - Daily Living Skills standard score from VABS-II
 - Socialization standard score from VABS-II
 - Motor Skills standard score from VABS-II
 - CSF GAG levels (log-scale)

- ABC standard score from VABS-II and
 - Communication standard score from VABS-II
 - Daily Living Skills standard score from VABS-II
 - Socialization standard score from VABS-II
 - Motor Skills standard score from VABS-II
 - CSF GAG levels (log-scale)

The sample SAS code is presented in [appendix 9.2.4](#).

Each pair of parameters will also be graphed as scatter plots.

6.7.5.3 Pharmacodynamic Outcome

- CSF and Urine GAG

CSF and Urine GAG observed levels, the change from baseline (Screening Visit) and percent change from baseline will be summarized by visit and treatment group. Mean observed values (\pm SE) and mean percent change compared to baseline will be plotted over time. The percent change compared to baseline is defined as the value at the corresponding visit multiplied by 100 and divided by the baseline value. One decimal will be kept for the percent compared to baseline. A van Elteren (vE) test, described in Section [6.7.5.1](#), will be used to compare treatment groups on the change from baseline and percent change to the Visit Week 52. Baseline age group (either <6 years or ≥ 6 years) and baseline GCA score (either ≤ 70 or >70) will be used as stratification factors in the van Elteren test. Analysis of both CSF total GAG and CSF heparan sulfate if available will be performed.

- CSF Albumin Concentration and Albumin Quotients

Albumin levels will be measured in samples of CSF and serum at Screening, Weeks 4, 16, 28, 40 and Visit Week 52. CSF albumin concentrations, CSF/serum albumin quotients ($Q_{\text{Alb}} = \text{CSF albumin/serum albumin}$) observed levels, the change and the percent change from baseline (Screening Visit) will be summarized by visit and treatment group to assess the blood-CSF barrier function ([Andersson et al. 1994](#)).

6.7.5.4 Rate of Change in GCA Score Analysis

The endpoint of rate of change analysis is the slope estimate from the linear regression of GCA score over time for each subject. Only subjects with at least 3 available GCA scores will be included in this analysis.

The slope estimate will be compared between the two treatment groups in HIT-HGT-094 using one-way ANOVA. The response variable in the ANOVA model is the weighted linear regression slope, where the weight is the inverse of one plus the sum of mean squared error (MSE) from the linear regression, i.e. $1/(1+\text{MSE})$.

The sample SAS code is presented in [appendix 9.2.5](#).

6.8 Analysis of Safety

All safety analyses will be descriptive, no statistical testing will be performed. All analyses described below refer to the pivotal study; safety data from the separate substudy will be listed due to small number of patients expected.

6.8.1 Adverse Events

AEs will be recorded throughout the study and at early termination. AEs and medical conditions will be coded using the MedDRA version 16.1 or higher.

Treatment emergent adverse event (TEAE) for control group is defined as all AEs occurring on or after the date of randomization and on or before the EOS visit. TEAE for IT-treatment group is defined as all AEs occurring on or after the date of the first IDDD implant surgery or first dose (whichever is earlier) and on or before the EOS visit (+ 30 days) or 2 weeks after the removal of the last IDDD (whichever is later). In general, an AE will be deemed TEAE if it cannot be definitively categorized by the available components (day, month, year) of the AE onset date with respect to the date of the randomization or the date of intervention (the date of the first IDDD implant surgery or first dose [whichever is earlier]). Summaries for the following AE categories will be presented:

1. Patients who experienced no AEs,
2. Patients who experienced at least one AE
3. Patients who discontinued due to an AE(s)
4. Patients who died
5. Patients who experienced at least one serious adverse event (SAE),
6. Patients who experienced at least one severe AE,
7. Patients who experienced at least one IV Elaprase infusion-related adverse event
8. Patients who experienced at least one idursulfase-IT-related adverse event
9. Patients who experienced at least one IT Treatment Regimen-related adverse event (i.e., related to one or more of the following: study drug, IDDD, device surgical procedure, and IT-administration process)
10. Patients who experienced at least one IDDD surgical procedure-related adverse event
11. Patients who experienced at least one IDDD-related adverse event
12. Patients who experienced at least one IT Administration Process-related adverse event

The AE categories 1-7 will be presented by each treatment group for patients in the Safety Population, while the AE categories 8-12 will be presented by each treatment group for patients in the Safety Population who received idursulfase-IT. IT administration process-related AEs will be summarized by IT administration method (LP or IDDD) as well. The AE categories 10-11 will be presented for patients in the Safety Population with the device implanted. The number and proportion of patients, and the number of corresponding AEs will be summarized in overall summary tables. In addition, the number and percentage of patients having any treatment-emergent AE within the AE categories 2, 5, 7-12 and the number of corresponding AEs will be displayed by MedDRA system organ class (SOC), preferred term (PT) and treatment group. The most common treatment-emergent AEs which happened in >10% patients in either treatment group will also be summarized by SOC and PT. The most common non-serious treatment-emergent AEs which happened in >5% patients in either treatment group will also be summarized by SOC and PT.

The number and percentage of patients having any treatment-emergent AE, and the number of corresponding AEs will be summarized by MedDRA SOC, PT and severity.

The number and percentage of patients having any treatment-emergent AE, and the number of corresponding AEs will be summarized by relationship to IV Elaprase, idursulfase-IT, and at least one IT treatment regimen, and summarized by MedDRA SOC and PT.

Furthermore, listings of all the AEs of categories 3-6 will be provided if applicable.

6.8.2 Clinical and CSF Laboratory Evaluations

Chemistry and hematology serum laboratory and CSF (standard chemistries, glucose, protein, and cell counts) values will be summarized in terms of the absolute value and change from baseline at each scheduled study visit by treatment group. The total protein, glucose and total nucleated cells will be plotted for individual patients over time by scheduled study visit.

All the laboratory (chemistry, hematology, urinalysis and CSF) values will be categorized as a patient having had (1) a Clinically Significant (CS) value at any time during the study post baseline, and (2) no CS values at any time during the study post baseline. The number and percentage of patients in each category will be presented by treatment dose and overall. Patients with a Normal or non-CS value for a particular laboratory parameter at baseline who experience a change to CS for that laboratory parameter post-baseline will be identified and listed separately along with the corresponding laboratory values over time. The baseline is defined as the Screening measurement. If it is not available, the Pre-Surgery Day 1 of Week 1 measurement will be used as baseline.

If a particular laboratory measurement has been either repeated or retested, then the repeated or retested measurement for that laboratory parameter, with respect to date/time, will be used in the statistical analysis unless this value is invalid/missing. The handling of repeated or retested laboratory measurements should only consider the specific laboratory measurement that was repeated or retested.

6.8.3 12-Lead ECG Evaluations

The 12-lead ECG parameters (heart rate [bpm], PR interval [msec], QRS interval [msec], QT interval [msec] and the corrected QT interval (QTc) [msec]) will be summarized in terms of absolute value and change from baseline by visit and treatment group. The QTc interval will be calculated using Bazett's formula as QT divided by the square root of RR interval. The number and percentage of patients with ECG abnormalities post baseline will be presented by treatment group.

6.8.4 Vital Signs

Vital signs (temperature [C], pulse [bpm], blood pressure [systolic and diastolic, mmHg], oxygen saturation, and respiration rate [per min]), will be measured at Screening, Visit Weeks 4, 16, 28, 40 and 52 for all patients. In addition, vital signs will be collected at the following time points (± 10 minutes) in association with IT administration of idursulfase-IT, defined as IT vital signs: within 15 minutes prior to IT administration, 30 minutes post end of IT administration, 60 minutes post end of IT administration, 120 minutes post end of IT administration, and 4 hours post end of IT administration.

Vital signs will be plotted for individual patients over the time by scheduled study visit. The IT vital signs will be plotted for individual patients over the time by scheduled study visit from the start of the IT injection for each IT injection.

6.8.5 Physical and Neurological Examinations

Clinically significant physical and neurological examination findings will be recorded and summarized either as part of the medical history or adverse event data. No additional summary or listing will be provided.

6.8.6 Other Observations Related to Safety

6.8.6.1 Immunogenicity

Anti-idursulfase antibody formation will be monitored throughout the study for both serum and CSF. The number and percentage of patients testing anti-idursulfase antibody positive and negative at each scheduled time point will be summarized by treatment group. Antibody titer values will be plotted for individual patients over time by scheduled study visit in patients with positive antibodies at or prior to each scheduled visit. The neutralizing antibodies (NAb) titer will be plotted similarly for patients who developed positive neutralizing antibodies at or prior to each scheduled visit.

6.8.6.2 CSF Enzyme Level

Cerebrospinal fluid will be obtained from patients in the treated arm of the pivotal study and in the substudy at Screening, during surgical implantation of the IDDD, prior to each intrathecal injection of investigational product, and at the EOS visit.

The observed idursulfase enzyme levels in CSF will be plotted for individual patients before each IT dose over time by scheduled study visit. The descriptive statistics will be presented with the graph. If a measurement is <LLOQ, then it will be replaced by the zero value.

6.8.6.3 Height and Weight and Head circumference

Height (cm), weight (kg), and head circumference (cm) observed values and the change from baseline (Screening Visit) will be summarized by scheduled study visit and treatment group.

6.8.6.4 Intracranial Pressure (ICP)

ICP measurement (cm of H₂O) observed values and change from baseline will be summarized by scheduled study visit (Screening and EOS) and treatment group.

6.8.6.5 Brain MRI

The MRI measurements at baseline and the Visit Week 52, and the change from baseline to Visit Week 52 will be summarized descriptively by treatment group for the following parameters:

- Brain Total Intracranial Volume (cm³)
- Brain Total Tissue Volume (cm³)
- Brain Total White Matter (cm³)

- Brain Total Gray Matter (cm³)
- Total CSF Volume (cm³)

6.8.6.6 Hearing Assessments

The number and percentage of patients in the hearing assessment categories (right ear hearing aid, left ear hearing aid, and sufficient hearing for study assessments) will be summarized by treatment group, and visit (Screening, Visit Weeks 16, 28, 40 and 52).

6.8.6.7 Device Performance

Device-related terminology

- Initial device implant: The first IDDD implant that a patient ever receives.
- Partial device revision: Surgical revision/replacement of one or more component(s) of the device; other component(s) of the original device remain implanted and are not affected (e.g., port revision).
- Full device revision: The device is removed (explanted) in its entirety and a completely new device is implanted, including complete device removal with immediate replacement and delayed device implant after previous removal.
- Complete device removal with immediate replacement: Immediately implant of a new device after a device had been completely removed.
- Delayed device implant after previous removal: Implant of a new device after a previous device had been completely removed without immediate replacement on a separate and earlier occasion.
- Complete device removal without replacement: All parts of the device (both port and catheter) are removed and there is no new device implant.
- Device adjustment: surgical procedure of the device that is not a port replacement or full device revision.
- Device malfunction: The device does not perform as intended, based on the description in the device's Instructions for Use, but does not require either a partial or full device revision. If at the time of a scheduled dosing it is not possible to administer a full medication dosage as per the standard administration steps detailed in the device's Instructions for Use due to a device related issue, the IDDD will be declared a device malfunction. A device malfunction could be resolved without the need for a full or partial surgical revision. Programmatically it is when the malfunction date is present and outcome of malfunction is resolved.
- Device failure: When the device irreversibly fails to perform as intended and cannot be corrected without a device surgical intervention of either a partial or full device revision or removal. The IDDD will be declared a device failure, starting from the date of the initial malfunction that persists leading to the above surgical intervention. IDDDs that are considered to be malfunctioning at the end of the study will be categorized as Failures. For programming purposes, a device failure is when the malfunction date is present and outcome of malfunction is either ongoing or device failure (Surgical Procedure). The date of the device failure is the date of the initial malfunction.

- IDDD longevity: Total number of IDDD failures and explant divided by the total time at risk across all subjects in weeks (sum of time to IDDD failure or the last injection if not failed at the end of the data cut or explant date, from initial implantation, delayed implantation, or revision for all IDDDs).

Device Performance Analyses

SOPH-A-PORT[®] safety and performance will be summarized for IDDD-implanted patients in Safety Population. The number and proportion of patients of the following categories and the corresponding event count and event percentage will be summarized:

- Patients who had the initial implant only (i.e., no additional surgeries)
- Patients who had any post-initial implantation device surgeries
- Patients who had any type of difficulty associated with the implant procedure (e.g., difficulty accessing spinal canal, etc)

IDDD-related surgical procedure details for initial implants, complete removal with immediate replacement, and delayed implantation across all IDDD implantations for all patients will be summarized by failed or not failed IDDDs and overall, including incision region (paramedian vs. other), identification of the catheter passer used (Phoenix Neuro vs. other), number of suture wings implanted, suture wing configuration, interspace for catheter insertion into lumbar spine (L1-2, L2-3, L3-4, L4-5, L5-S1), spinal vertebral level of catheter tip (cervical, thoracic, lumbar and sacral), use of a purse string suture, use of the guidewire, orientation of the Tuohy needle and clinical site.

The number and proportion of patients who had at least one abnormal IDDD radiological assessment finding and the number of abnormal findings from the IDDD radiological assessments will be also summarized by types of the abnormality.

The number and proportion of patients and IDDDs with total malfunctions (including failure and malfunctions) as well as the corresponding event numbers will be presented. The types of IDDD malfunctions and the reasons for IDDD failures reported by the site and the Adjudication Committee will be summarized at the patient, IDDD and event level.

The annual event rate of IDDD failures and malfunctions will be calculated for each patient and the descriptive statistics will be summarized. The overall IDDD failure rate and its 95% CI will be presented. The overall IDDD failure rate is calculated as the total number of IDDD failures for all patients divided by the total IDDD time at risk, which is defined as the total time to IDDD failure or the last injection if IDDD is not failed at the end of the study, from initial implantation, or full device revision for all IDDDs. The overall IDDD malfunction rate is defined in the same manner.

The IDDD longevity (time to failure in weeks) and time to first total malfunction, for all implanted IDDDs, will be analyzed using the KM method. A new port identifies a new IDDD, starting from the date of implantation (either an initial implantation, a partial or full revision, or a

delayed device implant after previous removal). The time to IDDD failure (weeks) or total malfunction will be obtained by subtracting the date of the IDDD kit implantation from the date of IDDD failure (i.e., the initial malfunction date that persists leading to surgical intervention) or first total malfunction plus 1, and divided by 7, one decimal will be kept. IDDDs which did not fail or malfunction will be censored at the last study drug injection date for each IDDD. The number of IDDDs at risk, the cumulative number of IDDDs failed, censored, and the cumulative probability of failure with its standard error will be summarized in a table at each event time.

The proportion of patients for whom a successful first injection of investigational drug product occurred will be reported among those for whom a first injection was attempted (i.e., those who had an apparently successful implantation and did not suffer a device removal or revision prior to first scheduled injection). An injection is defined as study drug administered using IDDD. A successful injection is defined as the completion of study drug injection via IDDD without a malfunction. The proportion of patients who had no unsuccessful injection attempts during the study will also be summarized. Injections those are not successful for patient-related reasons (e.g., patient uncooperative, competing medical issue) will not be included in the determination of the injection success rate. The detailed list of the “Other” malfunction types will be reviewed by the Shire physician to determine if the injection was not given for patient reasons. If the injection is not given for patient reasons then it would not be counted as an unsuccessful injection attempt.

Whether any device component was implanted will be summarized for all IDDD kits which were opened. For the implanted IDDDs, their status (i.e. removed or not) at the end of the study will be reported. If an IDDD was removed, whether the catheter was removed along with the port will also be summarized.

An IDDD timeline of events (in weeks from implantation) including the timing of device surgeries (device adjustment, partial or full revision, and complete device removal without replacement), and malfunction and failure by IDDDs will be plotted by patient sorted by the duration since the initial implantation, longest on top.

The number of surgeries per patient will be presented by a bar chart. The types of IDDD malfunctions and failure per patient, as well as the reasons for the failures for the IDDDs reported by the site and the Adjudication Committee will be presented similarly by a bar chart.

The time from initial implant surgery to first IDDD failure and the time to first malfunction will be analyzed using the Kaplan Meier (KM) method for each patient who is in the Safety Population with IDDD implanted. The time to first IDDD failure or malfunction will be obtained by subtracting the date of the first IDDD implantation for the patient from the date of first failure or malfunction plus 1, divided by 7, and one decimal will be kept. Patients without an IDDD failure or malfunction will be censored at their last study drug injection date. The 25th percentile (Q1), median and 75th percentile (Q3) of the time to first failure or malfunction distribution and the 95% confidence interval (CI) will be presented together with the KM plots. A by-patient listing of the device failure and malfunction data will be displayed.

The IDDD longevity (time to failure in weeks) and time to first total malfunction, for all implanted IDDDs, will be plotted using the KM method. The 25th percentile (Q1), median and 75th percentile (Q3) of the time to failure or total malfunction distribution and the 95% CI will also be presented together with the KM plots.

The time (weeks) from IDDD implantation to failure/continuing function will be plotted as a horizontal bar chart for each device sorted by duration of each device implantation (time from the implantation of the corresponding IDDD), longest on top.

6.8.7 Concomitant Medications

Concomitant medications for control group are defined as all medications taken on or after the date of randomization and on or before the EOS visit. Concomitant medications for IT-treatment group are defined as all medications taken on or after the date of the first IDDD implant surgery or first dose (whichever is earlier) and on or before the EOS visit (+ 30 days) or 2 weeks after the removal of the last IDDD (whichever is later).

Concomitant medications will be coded using the WHO Drug Dictionary Version September 2013 or higher. Concomitant medication use will be summarized by the Therapeutic Class and Preferred Term for each treatment group for the Safety Population.

6.8.8 Concomitant Therapies

Concomitant therapies for control group are defined as all therapies taken on or after the date of randomization and on or before the EOS visit. Concomitant therapies for IT-treatment group are defined as all therapies taken on or after the date of the first IDDD implant surgery or first dose (whichever is earlier) and on or before the EOS visit (+ 30 days) or 2 weeks after the removal of the last IDDD (whichever is later).

Concomitant therapies will be coded using the MedDRA Version 16.1 or higher and summarized by the System Organ Class and Preferred Term for each treatment group for the Safety Population.

6.9 Analysis of Pharmacokinetic Data

All pharmacokinetic analyses will be performed using the PK Analysis Population. Serum and CSF concentrations as well as pharmacokinetic parameters will be listed and summarized by treatment.

6.9.1 Concentration Data

Blood samples will be collected for determination of idursulfase serum concentration-time profiles and serum pharmacokinetic parameters after IT administration. Idursulfase concentrations in serum and CSF will be determined using a validated Enzyme-Linked Immunosorbent Assay (ELISA) method which was used for the previous idursulfase IT clinical studies. The idursulfase concentrations in serum will also be determined by a validated liquid chromatography tandem mass spectrometry (LC-MS/MS) method. The LC-MS/MS method is expected to be less susceptible to interference by anti-idursulfase antibodies.

6.9.2 Handling BLQ Values

The following procedures will be used for serum concentrations below the lower limit of quantification (LLOQ) (reported as BLQ):

- Serum samples that are BLQ are reported as zero on the data listings.
- Samples that are BLQ are treated as zero in the calculation of summary statistics (e.g., mean, SD, etc.) for the serum concentrations at individual time points.
- Mean concentrations are reported as zero if all values are BLQ, and no descriptive statistics are reported. If the calculated mean (\pm SD) concentration is less than the LLOQ, the value will be reported as calculated. The mean values derived using these conventions will be used to create the mean serum concentration versus time plots.
- For calculation of area under the serum concentration curve (AUC), BLQ values are set equal to zero in the dataset loaded into WinNonlin for pharmacokinetic analysis. WinNonlin uses the zero values that occur before the first time point with a concentration greater than LLOQ, but WinNonlin excludes the zero values from AUC calculation for all later time points.

6.9.3 Pharmacokinetic Parameters

The pharmacokinetic analysis will be conducted by the pharmacokinetics CRO for Clinical Pharmacology and Pharmacokinetics Department of Shire Pharmaceuticals using Phoenix WinNonlin version 6.2 or higher (Pharsight Corporation, Mountain View, California, USA).

Pharmacokinetic parameters will be determined from serum concentration-time data using noncompartmental methods and all calculations will be based on actual sampling times. Serum concentration vs. time will be plotted for each patient. Mean serum concentration vs. time curves will also be presented by dose (5 mg, 7.5 mg, and 10 mg) and visit (Week 4, Week 24, and Week 48).

The pharmacokinetic parameters to be determined will include, but not be limited to, the following:

C_{\max}	Maximum concentration occurring at t_{\max}
t_{\max}	Time of maximum observed concentration sampled during a dosing interval
$t_{1/2}$	Terminal half-life
$AUC_{0-\infty}$	Area under the curve extrapolated to infinity, calculated using the observed value of the last non-zero concentration
AUC_{0-t}	Area under the curve from the time of dosing to the last measurable concentration
CL/F	Total body clearance for extravascular administration divided by the fraction of dose absorbed
V_z/F	Volume of distribution associated with the terminal slope following

λ_z extravascular administration divided by the fraction of dose absorbed
First order rate constant associated with the terminal (log-linear) portion of the curve

CL/F and Vz/F will also be corrected for body weight.

Summary statistics (number of observations, mean, standard deviation, coefficient of variation, median, maximum, minimum, and geometric mean) will be determined for all pharmacokinetic parameters and presented by bioanalytical method (ELISA, LC-MS/MS), dose and visit. Serum and CSF concentrations of idursulfase at each nominal sampling time will also be summarized by bioanalytical method, dose and visit using descriptive statistics.

The association of the presence of anti-idursulfase antibodies and idursulfase concentration-time profiles and pharmacokinetic parameters will be evaluated, if applicable.

6.10 Drug-Drug and Drug-Disease Interactions

The study is not designed to evaluate drug-drug or drug-disease interactions and therefore the interactions will not be assessed in this study.

7. CHANGES IN THE CONDUCT OF THE STUDY OR PLANNED ANALYSES

7.1 Changes in the Conduct of the Study

Please see the amendment summaries for the HGT-HIT-094 Protocol.

7.2 Changes from the Analyses Planned in the Protocol

The changes from the analyses planned in the protocol included

- Changing in subgroup age cut off from <7 to <6
- Adding interaction terms in the MMRM modelling
- Adding a sensitivity analysis for the primary analysis using ITT population excluding subjects enrolled from the observational study HGT-HIT-090 who experienced a decrease in GCA score of ≥ 10 points over 12 months in study HGT-HIT-090 prior to being enrolled in study HGT-HIT-094 and a GCA score at HGT-HIT-094 Screening >85
- Adding MMRM modeling for T scores of six core subtests from DASII EY
- Adding subgroup analysis for the subgroup with baseline age < 6 years
- Adding subgroup analysis for the subgroup with baseline GCA ≤ 70 and baseline GCA > 70
- Adding subgroup analysis for the subgroup with baseline age < 55 months and baseline age ≥ 55
- Adding rate of change in GCA score analysis
- Changing non-parametric Wilcoxon Rank Sum test to van Elteren (vE) test

8. STATISTICAL/ANALYTIC ISSUES

8.1 Adjustment for Covariates

Analyses of the change from baseline in efficacy endpoints utilizing MMRM methodology will adjust for the baseline score of the parameter of interest as a continuous covariate in the model. As the randomization of the study is stratified by the baseline GCA classification factor (either ≤ 70 or > 70), MMRM analyses will also take into account the GCA strata as a classification factor. The baseline age group (either < 6 years or ≥ 6 years) is also a categorical covariate in the MMRM analysis.

8.2 Handling of Dropouts or Missing Data

For randomized patients who discontinue early from the study, their early EOS efficacy evaluations will be used for the next scheduled time point only if the EOS evaluation is performed within 45 days of the next scheduled visit. In general, no further imputation will be used for descriptive analyses, or for primary and secondary efficacy analyses utilizing MMRM methodology.

8.3 Interim Analyses and Data Monitoring

No formal interim analysis or interim statistical testing for treatment comparisons is planned.

Patient safety data only in this study will be monitored by an independent DMC until the last patient completes his last scheduled study visit/assessment. The DMC will be an external group overseeing the safety of the study treatment, including both the investigational product and the IDDD, and will operate according to a charter determining the scope of its activities and frequency of meetings.

8.4 Multi-center Studies

This is a multicenter study utilizing a central randomization. The randomization is not stratified by center. It is planned that the data from all centers that participate in this protocol will be combined so that an adequate number of patients will be available for analysis. Because of the potential for a relatively large number of centers, and small numbers of patients at some centers, no subset analyses by center are planned. No adjustment for center effect will be utilized in the statistical analyses.

8.5 Multiple Comparisons/Multiplicity

The family-wise type I error rate (FWER) for the statistical tests of the primary, key secondary and other selected secondary efficacy endpoints (specified below) will be controlled at 0.05. To strongly control the FWER at this level, a Gate-keeping approach will be utilized in which each family of statistical tests will be conducted in a sequential manner. The test for the primary endpoint will be conducted first at the 5% significance level and, if significant, the key secondary endpoint will be similarly tested at the 5% significance level. If these two tests are

both significant, test of the additional secondary efficacy endpoints specified below will be conducted using the Hommel closed testing procedure (1988) to control the FWER.

The hypotheses for the primary, key secondary and other selected secondary efficacy endpoints will therefore be grouped into 3 ordered Gatekeeper families defined as follows:

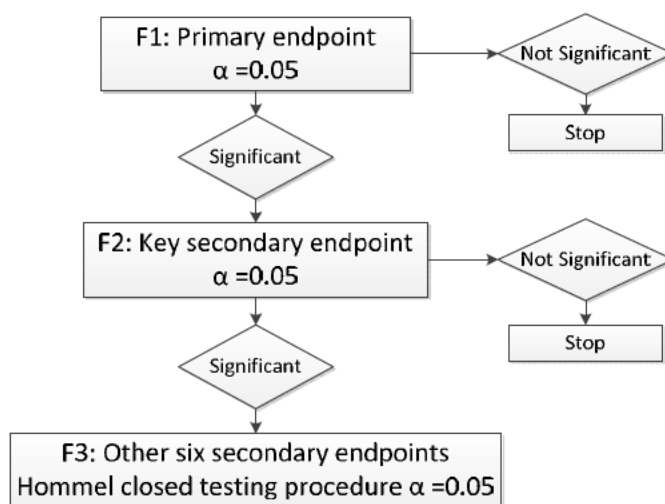
- Family 1 (F1): Hypothesis test for the primary endpoint
- Family 2 (F2): Hypothesis test for the key secondary endpoint
- Family 3 (F3): Hypothesis tests for the change from baseline to Visit Week 52 for the following 6 endpoints:
 - Verbal, Nonverbal and Spatial standard scores from the DAS-II;
 - Communication, Daily Living Skills and Socialization standard scores from the VABS-II;

The 3 sets of hypotheses in F1, F2 and F3 will be tested sequentially, with F1 and F2 serving as serial gatekeepers. The null hypothesis of no treatment difference for the primary endpoint will be tested first in F1. The statistical test for the key secondary endpoint in F2 may only be declared statistically significant if a significant result is also achieved for the primary efficacy endpoint. Since F1 and F2 each contain a single null hypothesis to be tested, each of the two tests will be conducted at the 5% significance level.

Further, only if the null hypotheses for both the primary and key secondary endpoints are rejected in F1 and F2, can any of the tests for the endpoints specified in F3 be declared statistically significant. If statistical significance is obtained for the hypothesis tests in F1 and F2, the Hommel (1988) closed testing procedure will be used to control the FWER within F3, with each hypothesis test in F3 equally weighted. In this case, SAS Proc MULTTEST will be used to calculate and report adjusted p-values for each endpoint in F3.

The sample SAS code is presented in the [appendix 9.2.6](#).

The multiplicity adjustment for the primary and secondary endpoints is shown in the following flowchart:



All other endpoints not included in F1, F2 or F3 above are considered supportive and any statistical tests comparing treatments will be made without adjustment for multiplicity. The resulting p-values from these supportive analyses will be interpreted descriptively as summarizing the weight of evidence for a treatment difference and may suggest avenues for further exploratory analyses or generate formal hypotheses to be tested in future trials.

8.6 Examination of Subgroups and Interactions

Please see section 6.7.4 for a description of subgroup analyses and tests for interaction.

8.7 Sensitivity Analyses

Sensitivity analyses for the primary and key secondary endpoints will be conducted to assess missing data assumptions in the MMRM analyses (Siddiqui et al. 2009). The sensitivity analyses below will use worst-observation-carried forward (WOCF) to impute missing data at the Visit Week 52 time point. Within a given patient, the worst (lowest) observed score prior to Visit Week 52 (including the baseline value) will be used for the imputation of their missing Visit Week 52 value.

The first sensitivity analysis will compare the treatment groups at the Visit Week 52 time point using an analysis of covariance model with WOCF (ANCOVA-WOCF). The ANCOVA-WOCF model will include fixed categorical effects for treatment, baseline GCA classification factor (either ≤ 70 vs. > 70), baseline age group (either < 6 years or ≥ 6 years), and the corresponding baseline score as a continuous covariate. The sample SAS code is presented in appendix 9.2.7. The treatment difference in least squares means, 95% confidence interval and p-value from the ANCOVA-WOCF model will be compared to the corresponding inferential statistics at Visit Week 52 from the primary or key secondary MMRM model.

The second sensitivity analysis, the van Elteren (vE) test with WOCF (vE-WOCF) will be used to compare treatment groups at Visit Week 52. This test is based on the ranked changes from baseline after imputation of the Visit Week 52 value using WOCF. Baseline age group (either < 6

years or ≥ 6 years) and baseline GCA score (either ≤ 70 or > 70) will be used as stratification factors in the van Elteren test. The sample SAS code is presented in [appendix 9.2.7](#).

If the primary or key secondary endpoint is missing in $> 15\%$ of patients in either treatment group, then an MNAR pattern mixture model (PMM) will be fit with control-based multiple imputation, according to the sequential imputation method described by Ratitch and O'Kelly (2011). Five hundred complete datasets will be imputed and the analysis of each imputed dataset will compare the treatment groups at the Visit Week 52 time point using an analysis of covariance (ANCOVA) model. The ANCOVA model will include fixed categorical effects for treatment, baseline GCA classification factor (either ≤ 70 vs. > 70), baseline age group (either < 6 years or ≥ 6 years), and the corresponding baseline score as a continuous covariate. The least squares mean difference estimates will be averaged and the associated SEs will be summarized based on within-imputation and between-imputation variance using the SAS MIANALYZE procedure to yield a final estimate with associated 95% CI and p-value. The sample SAS code is presented in [appendix 9.2.8](#).

If the MMRM model for the primary or key secondary endpoint fails to converge using an unstructured covariance matrix, an AR(1) covariance structure will be used. In this case, additional covariance structures will be tried as sensitivity analyses to assess the robustness of the results, including compound symmetric and Toeplitz structures.

If there are cases in which subjects receive a study treatment that differs from the treatment to which they were randomly assigned, MMRM, ANCOVA-WOCF and WRS-WOCF analyses of the primary and key secondary endpoints will be performed by treatment actually received.

8.8 Windowing Visits

Although there is a visit window of ± 7 days around the expected visit date, nominal visits will be used for the per-visit analyses. The data obtained outside the scheduled visits for certain procedures could be assigned to nominal visits, for instance, CSF GAGs or Neurodevelopment assessments. For subjects who withdraw from the study prematurely, if the early termination visit falls into the window of a scheduled visit as defined in the protocol, the early termination visit is also summarized for that scheduled visit, unless the scheduled visit already took place.

8.9 Data Listings

All data will be presented as SAS datasets in CDISC format. Unless specifically stated in above sections, no other by-patient data listings will be provided.

9. APPENDICES

9.1 Appendix I - List of Statistical Outputs

There are separate documents called “Table Shells” for the pivotal study and substudy, respectively, including Tables, Figures, and Listings.

9.2 Appendix II - Sample SAS code

Note that the dataset and variable names may be different in the real data.

9.2.1 Sample SAS code for the primary analysis

```
ods output estimates=est Tests3=FixF;
ods graphics on;
proc mixed data=adef data=adef covtest;
  class Trt VisitWk GCAclass Agegroup Genotype usubjid;
  model GCAchgBL = GCAclass Trt VisitWk Trt*Visitwk GCABL Agegroup Trt*GCAclass
  Trt*Agegroup GCAclass* Agegroup Genotype / ddfm=KR residual;
  repeated VisitWk / type=un subject=usubjid rcorr;
  estimate "Trt diff at Week 16" Trt 1 -1 Trt*VisitWk 1 0 0 0 -1 0 0 0 / cl;
  estimate "Trt diff at Week 28" Trt 1 -1 Trt*VisitWk 0 1 0 0 0 -1 0 0 / cl;
  estimate "Trt diff at Week 40" Trt 1 -1 Trt*VisitWk 0 0 1 0 0 0 -1 0 / cl;
  estimate "Trt diff at Week 52" Trt 1 -1 Trt*VisitWk 0 0 0 1 0 0 0 -1 / cl;
  lsmeans Trt*VisitWk / cl;
run;
```

9.2.2 Sample SAS code for the key secondary analysis

```
ods output estimates=est Tests3=FixF;
ods graphics on;
proc mixed data=adef data=adef covtest;
  class Trt VisitWk GCAclass Agegroup Genotype usubjid;
  model ABCchgBL = GCAclass Trt VisitWk Trt*Visitwk ABCBL Agegroup Trt*GCAclass
  Trt*Agegroup GCAclass* Agegroup Genotype / ddfm=KR residual;
  repeated VisitWk / type=un subject=usubjid rcorr;
  estimate "Trt diff at Week 16" Trt 1 -1 Trt*VisitWk 1 0 0 0 -1 0 0 0 / cl;
  estimate "Trt diff at Week 28" Trt 1 -1 Trt*VisitWk 0 1 0 0 0 -1 0 0 / cl;
  estimate "Trt diff at Week 40" Trt 1 -1 Trt*VisitWk 0 0 1 0 0 0 -1 0 / cl;
  estimate "Trt diff at Week 52" Trt 1 -1 Trt*VisitWk 0 0 0 1 0 0 0 -1 / cl;
  lsmeans Trt*VisitWk / cl;
run;
```

9.2.3 Sample SAS code for the EQ-3D analysis

```
proc npar1way data=adef ;
```



```
strata GCAclass Agegroup;  
class Trt;  
var VASchgBL;  
run;
```

9.2.4 Sample SAS code for the correlation analysis using repeated measures

```
ods output vcorr=vcorr;  
proc mixed;  
  class usubjid vtype replicate;  
  model response = vtype / ddfm=kr;  
  random vtype / type=un subject=usubjid v vcorr cl;  
  repeated vtype / type=un subject=replicate(usubjid) r rcorr;  
run;
```

usubjid corresponds to subject number; vtype refers to the two variables assessed, which are coded as 1 and 2; Response corresponds to the values of the two variables; and Replicate corresponds to the number of repeated measurements for each subject, which need not be the same.

9.2.5 Sample SAS code for rate of change in GCA score analysis

```
ods output ANOVA=ANOVA ParameterEstimates=Parms;  
proc reg data = subject_data ;  
  model GCA = VisitWk ;  
run ;Linear regression fit for individual subject data to calculate individual slope.
```

```
data weight;  
  set ANOVA(drop=Model DEPENDENT DF SS FVALUE PROBF);  
  weight=1/(1+MS);  
  if SOURCE='Error';  
run;
```

Two sample comparison:

```
proc glm;  
  class Trt;  
  model weighted_slope =trt;  
  lsmeans trt / pdiff CL;  
run;
```

where weighted slope=1/(1+MSE from linear regression fit)*slope.

9.2.6 Sample SAS code for the multiplicity analysis

```
proc multtest inpvalues=rawp hommel;  
run;
```

Dataset rawp contains the raw p-values from hypothesis tests for the change from baseline to Visit Week 52 using MMRM analysis for the following 6 endpoints:

- Verbal, Nonverbal and Spatial standard scores from the DAS-II;
- Communication, Daily Living Skills and Socialization standard scores from the VABS-II;

9.2.7 Sample SAS code for sensitivity analysis using WOCF

- ANCOVA-WOCF

```
proc mixed data=adefwocf;  
  class GCAclass Agegroup Trt;  
  model GCAchgBL = GCAclass Agegroup Trt GCABL AgeBL;  
  estimate "Diff. at month 13" Trt 1 -1/cl;  
  lsmeans Trt /cl;  
  ods output Estimates=E_ANCOVA;  
  where VisitWk=52;  
run;
```

- vE-WOCF;

```
proc npar1way data=adefwocf ;  
  strata Agegroup GCAclass;  
  class Trt;  
  var GCAchgBL;  
  where VisitWk=52;  
run;
```

9.2.8 Sample SAS code for sensitivity analysis using PMMs with control-based pattern imputation

A total of 1000 sets of posterior mean and co-variance estimates are extracted from the SAS MI procedure using the available non-missing placebo data. Five hundred of the posterior sets will be applied to the pooled active treatment group, while the other 500 will be applied to the placebo group. One set of imputations for all missing values will be generated based on each variation of posterior estimates. All 500 datasets for imputations within a treatment group will be ordered from 1 to 500 and combined between pooled active treatment group and placebo, for a total of 500 completely imputed datasets.

```
*Step 1;  
proc mi data=unimputed out=discard nimpute=1000 seed=609 noprint;  
  where Trt=0; *where 0 is placebo;  
  var GCABL GCA_wk16-GCA_wk52;  
  mcmc outest=posterior; *picks up the placebo posteriors;  
run;
```

```
*500 posteriors for the placebo, rest for active;  
data posteriors(type=est);  
  set posteriors;  
  if 1<=_imputation_<=500 then do; *assign placebo posteriors;  
    Trt=0;  
  end;  
  if 500<_imputation_<=1000 then do; *assign posteriors to active;  
    Trt=1;  
    _imputation_=_imputation_-500;  
  end;  
run;
```

```
proc sort data=unimputed;  
  by Trt;  
run;
```

```
proc mi data=unimputed out=imputed;  
  by Trt;  
  var GCABL GCA_wk16-GCA_wk52;  
  mcmc inest=posterior; *use the placebo posteriors;  
run;
```

```
*Step 2;  
data endpoint;  
  set imputed;  
  endpoint= GCA_wk52-GCABL;  
run;
```

```
proc sort data=endpoint;  
  by _imputation_;  
run;
```

```
proc mixed data= endpoint;  
  class GCAclass Agegroup Trt;  
  model endpoint = GCAclass Agegroup Trt GCABL AgeBL;  
  by imputation ;  
  estimate "Diff. at month 13" Trt 1 -1/cl;  
  lsmeans Trt /cl;
```

```
ods output Estimates=E_ANCOVA;  
run;
```

```
*Step 3;  
proc sort data= E_ANCOVA;  
  by label _imputation_ ;  
run;
```

```
proc mianalyze data= E_ANCOVA;  
  by label;  
  modeleffects estimate;  
  stderr stderr;  
run;
```

Repeat the process above for the key 2nd endpoint.

9.2.9 Sample SAS code for the subgroup analysis with baseline age < 6 years, baseline age < 55 months and baseline age ≥ 55 months

```
ods output estimates=est Tests3=FixF;  
ods graphics on;  
proc mixed data=adef_subgroup covtest;  
  class Trt VisitWk GCAclass Agegroup Genotype usubjid;  
  model GCAchgBL = GCAclass Trt VisitWk Trt*Visitwk Trt* GCAclass GCABL  
  Genotype/ ddfm=KR residual;  
  repeated VisitWk / type=un subject=usubjid rcorr;  
  estimate "Trt diff at Week 16" Trt 1 -1 Trt*VisitWk 1 0 0 0 -1 0 0 0 / cl;  
  estimate "Trt diff at Week 28" Trt 1 -1 Trt*VisitWk 0 1 0 0 0 -1 0 0 / cl;  
  estimate "Trt diff at Week 40" Trt 1 -1 Trt*VisitWk 0 0 1 0 0 0 -1 0 / cl;  
  estimate "Trt diff at Week 52" Trt 1 -1 Trt*VisitWk 0 0 0 1 0 0 0 -1 / cl;  
  lsmeans Trt*VisitWk / cl;  
run;
```

9.2.10 Sample SAS code for the subgroup analysis with baseline GCA classification factor

```
ods output estimates=est Tests3=FixF;  
ods graphics on;  
proc mixed data=adef_subgroup covtest;  
  class Trt VisitWk GCAclass Agegroup Genotype usubjid;  
  model GCAchgBL = Agegroup Trt VisitWk Trt*Visitwk Trt* Agegroup GCABL Genotype/  
  ddfm=KR residual;  
  repeated VisitWk / type=un subject=usubjid rcorr;
```

```
estimate "Trt diff at Week 16" Trt 1 -1 Trt*VisitWk 1 0 0 0 -1 0 0 0 / cl;  
estimate "Trt diff at Week 28" Trt 1 -1 Trt*VisitWk 0 1 0 0 0 -1 0 0 / cl;  
estimate "Trt diff at Week 40" Trt 1 -1 Trt*VisitWk 0 0 1 0 0 0 -1 0 / cl;  
estimate "Trt diff at Week 52" Trt 1 -1 Trt*VisitWk 0 0 0 1 0 0 0 -1 / cl;  
lsmeans Trt*VisitWk / cl;  
run;
```

9.2.11 Sample SAS code for subgroup analysis testing treatment interaction

```
ods output estimates=est Tests3=FixF;  
ods graphics on;  
proc mixed data=adefc covtest;  
  class Trt VisitWk GCAclass Agegroup Genotype usubjid;  
  model GCAchgBL = Trt VisitWk GCABL Trt*Visitwk Trt* GCABL Trt*Visitwk*GCABL  
  GCAclass Agegroup Genotype/ ddfm=kr solution;  
  repeated VisitWk / type=un subject=usubjid rcorr;  
  estimate "Trt diff at Week 16" GCABL 25th_of_BL_GCA_distribution Trt 1 -1 Trt*VisitWk  
  1 0 0 0 -1 0 0 0 / cl;  
  estimate "Trt diff at Week 28" GCABL 25th_of_BL_GCA_distribution Trt 1 -1 Trt*VisitWk  
  0 1 0 0 0 -1 0 0 / cl;  
  estimate "Trt diff at Week 40" GCABL 25th_of_BL_GCA_distribution Trt 1 -1 Trt*VisitWk  
  0 0 1 0 0 0 -1 0 / cl;  
  estimate "Trt diff at Week 52" GCABL 25th_of_BL_GCA_distribution Trt 1 -1 Trt*VisitWk  
  0 0 0 1 0 0 0 -1 / cl;  
  estimate "Trt diff at Week 16" GCABL 50th_of_BL_GCA_distribution Trt 1 -1 Trt*VisitWk  
  1 0 0 0 -1 0 0 0 / cl;  
  estimate "Trt diff at Week 28" GCABL 50th_of_BL_GCA_distribution Trt 1 -1 Trt*VisitWk  
  0 1 0 0 0 -1 0 0 / cl;  
  estimate "Trt diff at Week 40" GCABL 50th_of_BL_GCA_distribution Trt 1 -1 Trt*VisitWk  
  0 0 1 0 0 0 -1 0 / cl;  
  estimate "Trt diff at Week 52" GCABL 50th_of_BL_GCA_distribution Trt 1 -1 Trt*VisitWk  
  0 0 0 1 0 0 0 -1 / cl;  
  estimate "Trt diff at Week 16" GCABL 75th_of_BL_GCA_distribution Trt 1 -1 Trt*VisitWk  
  1 0 0 0 -1 0 0 0 / cl;  
  estimate "Trt diff at Week 28" GCABL 75th_of_BL_GCA_distribution Trt 1 -1 Trt*VisitWk  
  0 1 0 0 0 -1 0 0 / cl;  
  estimate "Trt diff at Week 40" GCABL 75th_of_BL_GCA_distribution Trt 1 -1 Trt*VisitWk  
  0 0 1 0 0 0 -1 0 / cl;  
  estimate "Trt diff at Week 52" GCABL 75th_of_BL_GCA_distribution Trt 1 -1 Trt*VisitWk  
  0 0 0 1 0 0 0 -1 / cl;  
  lsmeans "Trt lsmeans at Q1" trt01pn*avisitn / at blgcass=Q1 / cl;  
  lsmeans "Trt lsmeans at Q2" trt01pn*avisitn / at blgcass=Q2 / cl;  
  lsmeans "Trt lsmeans at Q3" trt01pn*avisitn / at blgcass=Q3 / cl;  
run;  
Repeat for change from baseline in ABC score
```

9.3 Appendix III - MMRM Simulation to Assess Analytical Methods

9.3.1 Introduction

The primary efficacy endpoint of study HGT-HIT-094 is the change from baseline in the Global Conceptual Ability (GCA) score at Visit Month 13 (Week 52), as obtained by DAS-II testing. Patients are assessed using the DAS-II at Baseline (Screening Visit), Month 4 (Week 16), Month 7 (Week 28), Month 10 (Week 40) and Month 13 (Week 52). A total of 42 patients are to be randomized using a 2:1 allocation ratio (28 active IT patients, 14 control patients) stratified by baseline GCA score (≤ 70 or > 70). It is expected that patients will have a GCA score of between 55 and 85 at baseline.

The proposed primary analysis of the primary endpoint will compare the treatment groups using a linear mixed model repeated measures (MMRM) analysis. The repeated measures are the change from baseline GCA scores obtained at the scheduled Visit Months 4, 7, 10 and 13, respectively. An alternative analytical approach is to use Analysis of Covariance at Month 13 with Last/Worst-Observation Carried-Forward to impute any missing data (ANCOVA-L/WOCF). This approach requires that, within a given patient, the last/worst (lowest) observed score prior to Month 13 be used for the imputation of their missing Month 13 value.

A series of data simulation studies were conducted to compare a linear mixed model repeated measures (MMRM) analysis with Analysis of Covariance using Last/Worst-Observation Carried-Forward (ANCOVA-L/WOCF). The methods and results from the simulation studies are described below.

9.3.2 Simulation Methods

The baseline GCA score was generated as a normal random variable with a mean of 70 and standard deviation of 15 in both arms. An accept-reject approach was used to restrict the range of values to 55 to 85 (about 2/3 of the sample met this criterion). The records were divided into sets of 42 patients representing a clinical trial and randomly assigned to active or control groups in a 2:1 ratio using a Bernoulli distribution with parameter, $p = 2/3$. In this manner, approximately 3600 replications of a clinical trial dataset, each with 42 patients, were simulated.

In each replication, multivariate normal (MVN) data was generated for the repeated measures on the change in GCA score using SAS Proc SIMNORM with the following assumptions. Based on our Phase I/II data in the target population of interest, and under the alternative hypothesis, we assumed a 1 unit decline per month in GCA score in the untreated control arm. As up to 2 months may elapse between the baseline assessment and the start of idursulfase-IT treatment, a mean decline of 2 points at Month 13 would be expected if IT treatment was effective in stabilizing the decline. Therefore, under the alternative hypothesis the change in GCA data were generated assuming the following row vectors of mean changes and variances for the change from baseline corresponding to Months 4, 7, 10 and 13 in the control and active IT arms respectively:

Control: $\mu_c = (-4, -7, -10, -13)$, $\sigma^2 = (3^2, 5^2, 8^2, 10^2)$

Active: $\mu_a = (-2, -2, -2, -2)$, $\sigma^2 = (3^2, 5^2, 8^2, 10^2)$

This leads to the following row vector of mean treatment differences for the change from baseline (i.e. $\mu_a - \mu_c$): $\delta_a = (2, 5, 8, 11)$. The expected standard error of the mean difference at Month 13 is therefore $10(\sqrt{1/14+1/28})=3.3$. Similarly, under the null hypothesis of no treatment difference, MVN outcomes were generated with mean treatment differences $\delta_0 = (0, 0, 0, 0)$ with the same variances.

For each simulation, the data were generated using a first order auto-regressive (AR(1)) covariance structure with ij^{th} matrix element $\sigma_i\sigma_j\rho^{|i-j|}$, where the correlation between adjacent time points was set to $\rho = 0.80$. The covariance and correlation matrix was thus specified as displayed in [Tables A1](#) and [A2](#) respectively.

Table A1: Theoretical AR(1) Covariance Matrix for Repeated Measures (in change from baseline)				
	Month 4	Month 7	Month 10	Month 13
Month 4	9.00	12.00	15.36	15.36
Month 7		25.00	32.00	32.00
Month 10			64.00	64.00
Month 13				100.00

Table A2: Theoretical AR(1) Correlation Matrix for Repeated Measures (in change from baseline)				
	Month 4	Month 7	Month 10	Month 13
Month 4	1.000	0.800	0.640	0.512
Month 7		1.000	0.800	0.640
Month 10			1.000	0.800
Month 13				1.000

The MMRM model included effects for treatment, month, treatment by month interaction and baseline GCA score as a continuous covariate. SAS Proc Mixed with restricted maximum likelihood estimation (REML) and an unstructured (UN) within-patient covariance structure was used. Note that a UN covariance structure was used in MMRM analyses even though an AR(1) structure was used to simulate the data, so as not to favor MMRM. The Kenward-Roger approximation was used to estimate denominator degrees of freedom. Similarly, the ANCOVA model for Month 13 contained effects for treatment and baseline GCA score as a continuous covariate. All scenarios were re-run adding a stratification factor for baseline GCA score (≤ 70 versus > 70) to each model. As the results for this set of simulations were nearly identical, results from the first set of simulations without the stratification factor are shown below. Power at Month 13, for the null and alternative hypothesis scenarios, was calculated for each analysis method using a 5% significance level.

9.3.3 Missing data

Monotone missing data with differential drop-out by treatment after month 4 was generated using a logistic regression model for the probability of drop-out. Simulations with data missing at random (MAR) and missing not at random (MNAR) were performed. Under MAR, conditional on the variables in the model, the missing data depends on the observed outcomes, however, under MNAR, missing data depends on both observed and unobserved outcomes. For MAR data, the probability of drop-out at a particular time point was dependent on treatment and the observed score at the previous visit. For MNAR scenarios, the probability of missing data also depended on the current visit score, which was unobserved in the subsequent analysis. In both cases, the logistic model was specified such that worse scores (greater declines) resulted in a higher probability of missing data. An overall drop-out rate of approximately 20% was assumed, with higher drop-out in the control versus the active arm (approximately 30% versus 15%) under the alternative hypothesis and equal drop-out under the null hypothesis. No imputation for missing data was employed for the MMRM analysis. LOCF was used to impute missing data for the ANCOVA analysis (ANCOVA-LOCF). Note that LOCF and WOCF are expected to be very similar methods in these simulations and the results should apply to both methods since a mean monotonic decline in the control arm and a constant value after Month 4 in the active arm was assumed under the alternative hypothesis. A mean monotonic decline in both arms is assumed under the null hypothesis.

9.3.4 Results

Data generated under the Alternative Hypothesis.

9.3.4.1 Complete Data

Complete MVN data were initially generated as described above. [Tables A3](#) and [A4](#) display the resulting covariance and correlation matrices respectively, which are very consistent with the values of the parameters specified above in [Tables A1](#) and [A2](#).

	Month 4	Month 7	Month 10	Month 13
Month 4	8.966	11.943	15.295	15.200
Month 7		24.915	31.887	31.699
Month 10			63.809	63.385
Month 13				99.114

	Month 4	Month 7	Month 10	Month 13
Month 4	1.000	0.799	0.639	0.510
Month 7		1.000	0.799	0.639
Month 10			1.000	0.799
Month 13				1.000

The MMRM and ANCOVA models were fit to the complete data. [Table A5](#) displays the average observed and least squares means and standard errors from MMRM and ANCOVA across the approximately 3600 simulated trials for the mean treatment differences at each month. MMRM returns the expected mean difference estimates (i.e. estimates of $\delta_a = (2, 5, 8, 11)$). Observed and MMRM estimated standard errors were also similar at each Month. Estimates were nearly identical for MMRM and ANCOVA at Month 13. Based on a 5% significance level, the power for the MMRM and ANCOVA tests at Month 13 were 89.0% and 88.4% respectively. Of note, the MMRM estimated unstructured covariance was very close to the true values and there were no issues of model non-convergence despite the relatively small sample size in each simulated trial.

Table A5: Observed and Estimated Mean Treatment Differences (in change from baseline): Complete Data			
MONTH	Observed Mean (SE)	MMRM LSMean (SE)	ANCOVA-LOCF LSMean (SE)
4	1.981 (1.004)	1.983 (1.013)	
7	4.998 (1.696)	5.000 (1.701)	
10	8.034 (2.734)	8.037 (2.739)	
13	11.011 (3.377)	11.013 (3.382)	11.016 (3.417)

9.3.4.2 MAR Missing Data

[Table A6](#) presents the average cumulative percent of non-missing data at each month within each treatment arm (averaged across 3580 simulated trials). At Month 13, we have approximately 15% and 27% MAR missing data in active and control arms respectively.

Table A6: Average Percent of Non-Missing Data at each Month		
Treatment	Month	Mean Percent
Active	4	100.0%
	7	94.9%
	10	90.1%
	13	85.3%
Control	4	100.0%
	7	91.2%
	10	82.1%
	13	72.7%

In [Table A7](#) we see the average observed and estimated means and standard errors for treatment differences. MMRM returns means that are close to the true means (i.e. estimates of $\delta_a = (2, 5, 8, 11)$) with slightly larger standard errors (SE), while ANCOVA-LOCF is biased downward at Month 13 both in terms of the estimated mean and standard error. The average ANCOVA-LOCF bias in the mean was about -15%. Based on a 5% significance level, the power for the test at Month 13 for the MMRM model was 83.7%, while the power for ANCOVA-LOCF was 81.6%. Power was fairly similar for both analytical methods because the ANCOVA-LOCF underestimated both the mean treatment difference and its standard error.

Month	Observed Mean (SE)	Observed-LOCF Mean (SE)	MMRM Estimated LSMean (SE)	ANCOVA-LOCF LSMean (SE)
4	1.981 (1.004)	1.981 (1.004)	1.983 (1.013)	
7	4.972 (1.772)	4.726 (1.671)	4.993 (1.738)	
10	7.940 (2.971)	7.231 (2.617)	8.067 (2.873)	
13	10.733 (3.901)	9.386 (3.199)	11.051 (3.722)	9.389 (3.241)

9.3.4.3 MNAR Missing Data

Table A8 presents the average cumulative percent of non-missing data at each month within each treatment arm averaged across the simulated trials. At Month 13, we have approximately 14% and 29% MNAR missing data in active and control arms respectively.

Treatment	Month	Mean Percent
Active	4	100.0%
	7	95.7%
	10	91.1%
	13	86.3%
Control	4	100.0%
	7	92.4%
	10	82.5%
	13	70.8%

In Table A9 we see the average observed and estimated means and standard errors for treatment differences. Both MMRM and ANCOVA-LOCF return estimated means that are biased downward relative to the true means (i.e. estimates of $\delta_a = (2, 5, 8, 11)$), however, ANCOVA-LOCF is more biased at Month 13 both in terms of the estimated mean and standard error. The average bias in the mean at month 13 was about -3% for MMRM, but it was about -17% for ANCOVA-LOCF. Based on a 5% significance level, the power for the test at Month 13 for the MMRM model was 82.4%, while the power for ANCOVA-LOCF was 81.1%.

Month	Observed Mean (SE)	Observed-LOCF Mean (SE)	MMRM Estimated LSMean (SE)	ANCOVA-LOCF LSMean (SE)
4	1.988 (0.996)	1.988 (0.996)	1.991 (1.011)	
7	4.897 (1.693)	4.709 (1.605)	4.945 (1.668)	
10	7.611 (2.821)	7.091 (2.478)	7.881 (2.737)	
13	9.922 (3.725)	9.101 (3.017)	10.682 (3.605)	9.106 (3.055)

Table A10 summarizes the statistical power under the alternative hypothesis for complete, MAR and MNAR data scenarios for the MMRM and ANCOVA-LOCF models.

Missing Data Type	MMRM	ANCOVA-LOCF
Complete	89.0%	88.4%
MAR Data	83.7%	81.6%
MNAR Data	82.4%	81.1%

9.3.4.4 Data generated under the Null Hypothesis

Complete, MAR and MNAR missing data were generated assuming the mean treatment difference at each month was zero ($\delta_\theta = (0, 0, 0, 0)$) with the same variances as described above. The resulting covariance and correlation matrices for the complete data was nearly identical to that shown in [Tables A1](#) and [A2](#). The MMRM model was fit to the complete data. [Table A11](#) displays the average observed and least squares means and standard errors from MMRM and ANCOVA across the approximately 3600 simulated trials for the mean treatment differences at each month. MMRM returns the expected mean difference estimates (δ_θ). Observed and MMRM estimated standard errors were also similar at each Month. Estimates from MMRM are nearly identical to ANCOVA at Month 13. The type I error rate associated with the MMRM and ANCOVA test at Month 13 was 4.80% and 4.69% respectively, both close to the nominal 5%.

MONTH	Observed Mean (SE)	MMRM LSMean (SE)	ANCOVA LSMean (SE)
4	-0.019 (1.004)	-0.017 (1.013)	
7	-0.002 (1.696)	0.000 (1.701)	
10	0.034 (2.734)	0.037 (2.739)	
13	0.010 (3.377)	0.012 (3.382)	0.015 (3.417)

For null missing data scenarios, drop-out was approximately 20% in each treatment arm. Type I error rates were similar for MMRM and ANCOVA-LOCF at Month 13 and close to the nominal 5% level for the MAR data scenario as shown in [Table A12](#). For MNAR data, MMRM better controlled the type I error rate.

Missing Data Type	MMRM	ANCOVA-LOCF
Complete Data	4.80%	4.69%
MAR Data	4.94%	4.89%
MNAR Data	4.53%	5.29%

9.3.5 Conclusions

The following conclusions can be made from these simulation results:

- In performing the MMRM analysis using an unstructured covariance structure across all sets of simulations, we encountered no model convergence problems despite the relatively limited sample size
- In the absence of missing data, MMRM returned the correct means and variances at each Month for data generated under both the null and alternative hypotheses and the estimates at Month 13 were nearly identical to ANCOVA
- In simulations with 20% MAR data (approximately 15% and 30% missing data in active and control arms respectively), the MMRM estimates of the mean (SE) treatment differences remained correct, while ANCOVA-LOCF produced estimates that were biased downward at the primary Month 13 time point
- In simulations with 20% MNAR data (approximately 15% and 30% missing data in active and control arms respectively), both the MMRM and ANCOVA-LOCF estimates of the mean (SE) treatment differences were biased downward, however the average ANCOVA-LOCF Month 13 estimate was substantially more biased, underestimating the true mean value by more than 15%
- For all missing data scenarios under the alternative hypothesis, statistical power was similar for MMRM and ANCOVA-LOCF (approximately 80-89%). This was true for MAR and MNAR data scenarios in part because ANCOVA-LOCF tended to underestimate both the true mean treatment difference and its standard error. For all missing data scenarios under the null hypothesis, type I error rates were similar for the two methods, and close to the expected 5% level, although MMRM demonstrated better control of type I error under MNAR data.

9.3.6 Discussion

In this simulation study, MMRM and ANCOVA-LOCF were compared under two missing data scenarios: data missing at random (MAR) and data missing not at random (MNAR). In the scenarios considered here, LOCF and WOCF imputation methods would be expected to yield very similar results since either stable or monotonic declines in mean change in GCA score were assumed in the simulations. The assumption of steady decline in GCA score in the untreated control arm is justified because in the target patient population of interest, cognitive impairment is expected to progressively worsen at a rate of about 1 standard score point per month in the absence of an effective treatment.

Single imputation methods such as LOCF or WOCF assume that the data are missing completely at random (MCAR) and that the patient's last or worst observed score provides a valid estimate for their missing Month 13 value. MMRM analysis on the other hand requires the less restrictive assumption of data missing at random (MAR) and does not require explicit imputation of missing values. It is therefore not surprising that MMRM outperformed ANCOVA-LOCF in terms of bias under the simulated MAR data. More interesting is that under the simulated MNAR data, which violates the assumptions of both methods, MMRM was proven to be more robust than ANCOVA-LOCF, and therefore more appropriate as a primary method of analysis.

MMRM was also shown to have slightly greater power under alternative scenarios and better control of type I error in null scenarios under MNAR. Although these differences were small, they would likely become greater as the rate of missing data increased beyond the 20% considered here. Across all scenarios, there were no model convergence problems with the MMRM analysis, which was generally more reliable than ANCOVA-LOCF.

The findings in this simulation study were very similar to other published studies. Mallinckrodt et al. (2001a; 2001b) compared MMRM to ANOVA-LOCF in a series of simulation studies with monotone MNAR missing data. In comparisons of complete data, both methods yielded identical results. In simulations with an overall 40% rate of missing data and a true treatment effect, MMRM estimates were closer to the true value in every scenario assessed. Standard errors from MMRM were accurate, while ANOVA-LOCF underestimated uncertainty. In similar simulations with no treatment effect, the average type I error rate across scenarios using a significance level of 5% for MMRM with unstructured covariance was 5.9% compared with 10.4% for ANOVA-LOCF.

In extensive simulation studies of MCAR, MAR and MNAR data, Siddiqui et al. (2009) compared the performance of ANCOVA-LOCF with MMRM to assess empirical bias and type I error. The simulation studies demonstrated that LOCF can lead to substantial bias in estimators of treatment effect and can greatly inflate type I error rates, whereas MMRM analysis on the available data led to estimators with comparatively small bias and type I error rates at close to the nominal level. In a sensitivity analysis of 48 clinical trial datasets obtained from 25 New drug Application (NDA) submissions of neurological and psychiatric drug products, Siddiqui et al. found that MMRM analysis appeared to be a superior approach in controlling type I error rates and minimizing bias compared to ANCOVA-LOCF. In exploratory analyses of the NDA datasets, no clear evidence of the presence of MNAR missingness was found, indicating that the MMRM MAR assumption is a robust and often reasonable assumption for clinical trial data.

Based on these simulation results, and the published findings in the literature, Shire believes MMRM is an appropriate statistical method to evaluate the efficacy of HGT-2310 in the primary analysis of the primary endpoint in study HGT-HIT-094. Sensitivity analyses, including ANCOVA-WOCF, will also be conducted to assess the validity of the MAR assumption in the primary MMRM analysis.

10. REFERENCES

- Andersson M, Alvarez- Cermeño J , Bernadi G , et al. Cerebrospinal fluid in the diagnosis of multiple sclerosis: a consensus report . *J Neurol Neurosurg Psychiatry* 1994 ; 57 : 897 – 902.
- Hamlett A, Ryan L, Serrano-Trespacios P, Wolfinger R. Mixed models for assessing correlation in the presence of replication. *J Air Waste Manag Assoc* (1995) 2003;53:442-50.
- Hommel, G., (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika* 75, 383-386.
- Malinckrodt C, Lane P, Schnell D, Peng Y, Mancuso J. Recommendations for the Primary Analysis of Continuous Endpoints in Longitudinal Clinical Trials. *Drug Inf J* 2008;42:303-19.
- Mallinckrodt, C.H., Clark, W.S., David, S.R. Accounting for drop-out bias using mixed effects models. *J Biopharm Stat* 2001;11:9-21
- Mallinckrodt, C.H., Clark, W.S., David, S.R. Type I error rates from mixed effects model repeated measures versus fixed effects ANOVA with missing values imputed via last observation carried forward. *Drug Inf J* 2001;35:1215-1225
- Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. *Ann Math Stat* 1947; 18:50–60.
- Meikle PJ, Hopwood JJ, Clague AE, Carey WF. Prevalence of lysosomal storage disorders. *JAMA* 1999;281:249-54.
- Qu Y, Zhao YD, Rahardja D (2008) Wilcoxon-Mann-Whitney test: stratify or not? *J Biopharm Stat* 18: 1103-1111.
- Ratitch, B. and O’Kelly, M. Implementation of Pattern-Mixture Models Using Standard SAS/STAT Procedures. *PharmaSUG*, 2011.
- Siddiqui O., Hung H.M., O’Neil R. MMRM versus LOCF: A Comprehensive Comparison based on Simulation study and 25 NDA Datasets. *J Biopharm Stat* 2009;19:227-246.
- van Elteren PH. On the combination of independent two sample tests of Wilcoxon. *Bulletin of the International Statistical Institute* 1960; 37:351–361.
- Wilcoxon F. Individual comparisons by ranking methods. *Biometrics* 1945; 1:80–83.
- Young ID, Harper PS. Incidence of Hunter's syndrome. *Hum Genet* 1982;60:391-2.