Official Title of Study:

A PHASE 2, DOUBLE-BLIND, RANDOMIZED, PLACEBO-CONTROLLED, MULTICENTER STUDY TO DETERMINE THE EFFICACY AND SAFETY OF LUSPATERCEPT (ACE-536) VERSUS PLACEBO IN ADULTS WITH NON-TRANSFUSION DEPENDENT BETA (β)-THALASSEMIA

(The BEYOND™ Study)

NCT Number: NCT03342404

Document Date (Date in which document was last revised): Nov 5, 2020

STATISTICAL ANALYSIS PLAN

A PHASE 2, DOUBLE-BLIND, RANDOMIZED, PLACEBO-CONTROLLED, MULTICENTER STUDY TO DETERMINE THE EFFICACY AND SAFETY OF LUSPATERCEPT (ACE-536) VERSUS PLACEBO IN ADULTS WITH NON-TRANSFUSION DEPENDENT BETA (β)-THALASSEMIA

(The BEYOND™ Study)

STUDY DRUG: Luspatercept (ACE-536)

PROTOCOL NUMBER: ACE-536-B-THAL-002

DATE FINAL: 5 NOV 2020

Prepared by:

Celgene, a wholly owned subsidiary of

Bristol-Myers Squibb Company

[redacted]

CONFIDENTIAL

The information contained in this document is regarded as confidential and, except to the extent necessary to obtain informed consent, may not be disclosed to another party unless such disclosure is required by law or regulations. Persons to whom the information is disclosed must be informed that the information is confidential and may not be further disclosed by them.

TABLE OF CONTENTS

SIGNATURE PAGE

STATISTICAL ANALYSIS PLAN (SAP) AND SAP AMENDMENT APPROVAL SIGNATURE PAGE

SAP TITLE	ACE-536-B-THAL-002 Statistical Analysis Plan
SAP VERSION, DATE	Final Version 1.0 5 November 2020
[redacted]	[redacted]
PROTOCOL TITLE	A Phase 2, Double-Blind, Randomized, Placebo-Controlled, Multicenter Study To Determine The Efficacy And Safety of Luspatercept (ACE-536) Versus Placebo In Adults With Non Transfusion Dependent Beta (β)-Thalassemia
INVESTIGATIONAL PRODUCT	Luspatercept (ACE-536)
PROTOCOL NUMBER	ACE-536-B-THAL-002
PROTOCOL VERSION, DATE	12-JUN-2020
SIGNATURE STATEMENT	By my signature, I indicate I have reviewed this SAP and find its contents to be acceptable.

Statistical Therapeutic Area Head

Signature

Printed Name [redacted]

Lead Clinical Research Physician / Clinical Research Physician

Signature

Printed Name [redacted]

Lead Product Safety Physician

Signature

Printed Name [redacted]

1. LIST OF ABBREVIATIONS

6MWT	6-Minute Walk Test
ADA	Anti-drug antibody
ALT	Alanine aminotransferase (SGPT)
ANC	Absolute neutrophil count
ANCOVA	Analysis of covariance
AST	Aspartate aminotransferase (SGOT)
ATC	Anatomical therapeutic chemical
AUC	Area under the curve
BMD	Bone mineral density
BMI	Body mass index
BP	Bodily Pain
BSA	Body surface area
BUN	Blood urea nitrogen
CI	Confidence interval
CMH	Cochran-Mantel-Haenszel
CTCAE	Common Terminology Criteria for Adverse Events
DBP	Diastolic blood pressure
DBPT	Double-Blind Treatment Period
DMC	Data Monitoring Committee
DXA	Dual energy x-ray absorptiometry
Dw	Dry weight
ECG	Electrocardiogram
ECHO	Echocardiography
ECOG	Eastern Cooperative Oncology Group
eCRF	Electronic case report form
FACIT-F	Functional Assessment of Chronic Illness Therapy-Fatigue

FS	Fatigue Subscale
GDF	Growth differentiation factor
GH	General Health
Hb	Hemoglobin
HbF	Fetal hemoglobin
HRQoL	Health-related QoL
HRU	Healthcare Resource Utilization
ICF	Informed consent document
ICT	Iron chelation therapy
IP	Investigational product
IRT	Interactive Response Technology
ITT	Intent-to-treat
IVRS	Integrated voice response system
IWRS	Integrated Web Response System
LDH	Lactic dehydrogenase
LIC	Liver iron concentration
LLN	Lower limit of normal
LVEF	Left ventricular ejection fraction
MAA	Marketing authorization application
MCH	Mean corpuscular hemoglobin
MCHC	Mean corpuscular hemoglobin concentration
MCS	Mental component summary
MCV	Mean corpuscular volume
MedDRA	Medical Dictionary for Regulatory Activities
MH	Mental Health
MRI	Magnetic resonance imaging
MUGA	Multi Gated Acquisition Scan
NA	Not applicable
NCI	National cancer institute

NTDT	Non-transfusion dependent β-thalassemia
NTDT-PRO	Non transfusion dependent β-thalassemia-patient reported outcome
OLP	Open Label Phase
OR	Odds Ratio
PCS	Physical component summary
PF	Physical functioning
PK	Pharmacokinetic
PGI-C	Patient Global Impression of Change
PPS	Per Protocol Set
PTFP	Post-Treatment Follow-up Period
QoL	Quality of life
RBC	Red blood cell
RD	Responder Definition
RDW	Red blood cell distribution width
RE	Role-Emotional
RP	Role-Physical
SAP	Statistical analysis plan
SBP	Systolic blood pressure
SC	Subcutaneous
SD	Standard deviation
SE	Standard error
SF	Social functioning
SF-36	Medical Outcomes Study 36-Item Short Form
SI	Systeme Internationale
SGOT	Serum glutamic oxaloacetic transaminase (AST)
SGPT	Serum glutamic pyruvic transaminase (ALT)
SoB	Shortness of breath (domain of NTDT-PRO)
SOC	System and organ class
PT	Preferred term

TEAE	Treatment emergente adverse event
TRV	Tricuspid regurgitant velocity
T/W	Tiredness and weakness (domain of NTDT-PRO)
ULN	Upper limit of normal
VT	Vitality
WHO-DD	World Health Organization Drug Dictionary

[redacted]

3. STUDY OBJECTIVES

3.1. Primary Objective

The primary objective is to evaluate the effect of luspatercept versus placebo on anemia, as measured by mean hemoglobin concentration in the absence of transfusions over continuous 12-week intervals, from Week 13 to Week 24, compared to baseline.

3.2. Secondary Objectives

The secondary objectives are:

- To evaluate the effect of luspatercept versus placebo on β-thalassemia-related symptoms, as measured by non-transfusion-dependent β-thalassemia patient-reported outcome (NTDT-PRO) over continuous 12-week intervals (Weeks 13 to 24, Weeks 37 to 48) compared to baseline
- To evaluate the effect of luspatercept versus placebo on functional and health-related quality of life (QoL) as measured by Medical Outcomes Study 36-Item Short Form (SF-36) and Functional Assessment of Chronic Illness Therapy-Fatigue (FACIT-F) questionnaires
- To evaluate the long-term effect of luspatercept versus placebo on anemia, as measured by mean hemoglobin concentration in the absence of transfusions over a continuous 12-week interval from Week 37 to Week 48, compared to baseline
- To evaluate the effect of luspatercept versus placebo on iron overload, as measured by liver iron concentration (LIC) and iron chelation therapy (ICT) daily dose
- To evaluate the effect of luspatercept versus placebo on iron overload, as measured by serum ferritin
- To evaluate the duration of erythroid response
- To evaluate the effect of luspatercept versus placebo on physical activity measured by 6-minute walk (6MWT)

The safety and PK objectives are:

- To evaluate safety and tolerability of luspatercept, including immunogenicity
- To evaluate population pharmacokinetics (PK) of luspatercept in subjects with β-thalassemia

[redacted]

4. INVESTIGATIONAL PLAN

4.1. Overall Study Design and Plan

This is a Phase 2, double-blind, randomized, placebo-controlled, multicenter study to determine the efficacy and safety of luspatercept (ACE-536) versus placebo in adults with non-transfusion dependent beta (β)-thalassemia. The study is divided into the Screening Period, Double-blind Treatment Period (DBTP), Open-label Phase (OLP) and Post-treatment Follow-up Period (PTFP). The overall study design is described in Figure 1. It is planned to randomize approximately 150 subjects at a 2:1 ratio of luspatercept versus placebo.

4.1.1. Screening Period

Upon giving written informed consent, the subject will enter the Screening Period to determine eligibility. The Screening Period is to last up to 4 weeks with an administrative window of + 3 days to allow all laboratory results to be available at site for evaluating the subject's eligibility. Re-screening is allowed and a new subject ID number will be assigned. Subjects will be stratified at the time of randomization based on baseline hemoglobin level and baseline NTDT-PRO tiredness and weakness (T/W) score.

4.1.2. Double-blind Treatment Period

Subjects will enter the Double-blind Treatment Period (DBTP) once they have completed the required assessments in the Screening Period and been randomized via the Interactive Response Technology (IRT) system. The DBTP will begin on Dose 1 Day 1 and end when all subjects have completed 48 weeks of treatment or have discontinued earlier. At that time, the study will be unblinded.

Subject randomization will occur via the IRT system and Dose 1 Day 1 should be scheduled within 3 days of randomization (can be on the same day as randomization). Eligible subjects will be randomized at a ratio of 2:1, luspatercept versus placebo, at a starting dose level of 1.0 mg/kg administered by subcutaneous (SC) injection once every 3 weeks. The maximum total dose per administration is 120 mg.

Best supportive care is allowed in both the luspatercept and placebo groups. This will include RBC transfusions, iron-chelating agents, use of antibiotic therapy, antiviral and antifungal therapy, nutritional support as needed, and other medications that are not prohibited, thus minimizing the safety risk to patients.

After the study is unblinded, and after DMC's recommendation:

- Subjects who received placebo and have been assessed as per protocol up to 48 weeks after the first dose of investigational product (IP) (even if the IP is discontinued before completing 48 weeks of treatment), may access the Open Label Phase (OLP) to receive luspatercept for a maximum 15 months before moving to the rollover protocol for longer treatment.

- Subjects receiving luspatercept may continue their treatment in the OLP for maximum 15 months before moving to the rollover protocol for longer treatment (i.e. 5 years from Dose 1, or until treatment discontinuation, whichever occurs later).
- Subjects who received luspatercept and discontinued the IP in the DBTP may continue and/or complete the long-term follow-up of 5 years from first dose of IP, or 3 years from last dose (whichever occurs later), to complete the Post-Treatment Follow-Up Period (PTFP) under the rollover protocol in this study until the End of Trial.

Subjects may be discontinued from treatment and/or the study decided by the treating physician.

4.1.3. Open-Label Phase

Open-label Phase (OLP) will begin after the study unblinding, and after DMC's determination of risk/benefit analysis and recommendation. The start of this OLP will be determined by the availability of primary analysis data that justify the use of luspatercept in an OLP, which will be reviewed by the independent external DMC. After DMC review of safety and efficacy, DMC will determine if the use of luspatercept in subjects previously randomized to receive placebo in this OLP is safe and recommended, and if subjects already on luspatercept can continue to be treated at their current dose level (best supportive care is allowed). In the OLP, subjects may receive luspatercept for maximum 15 months or discontinue early.

Access to the OLP:

- Subjects initially assigned to placebo in the DBTP can enter the OLP only if the DMC allows it, as outlined above, and if:
  - they are still receiving placebo at the time of unblinding, or
  - they discontinued the IP before the unblinding, but they continued their participation in the PTFP until the unblinding and complied with the PTFP assessments, and they still fulfill the following selected eligibility criteria prior to Dose 1 Day 1, as per PI assessment using central lab data.
    - Inclusion criteria: numbers 8-10 (Refer to Section 4.2 of Protocol)
    - Exclusion criteria: numbers 1-8, 10, 12-15, 17, 18 and 20 (Refer to Section 4.3 of Protocol)
- Subjects initially assigned to luspatercept in the DBTP can enter the OLP if they are still receiving luspatercept at the time of unblinding, and may continue their treatment in the rollover protocol, after completion of the OLP.

Subjects who discontinue from the study before the unblinding and without completing the PTFP period are not allowed to re-enter this study and access luspatercept in the OLP.

4.1.4. Post-treatment Follow-up Period

Subjects who discontinue treatment with the IP in the DBTP or OLP, regardless of reason, will enter the PTFP and may continue or complete the long-term follow-up of 5 years from first dose

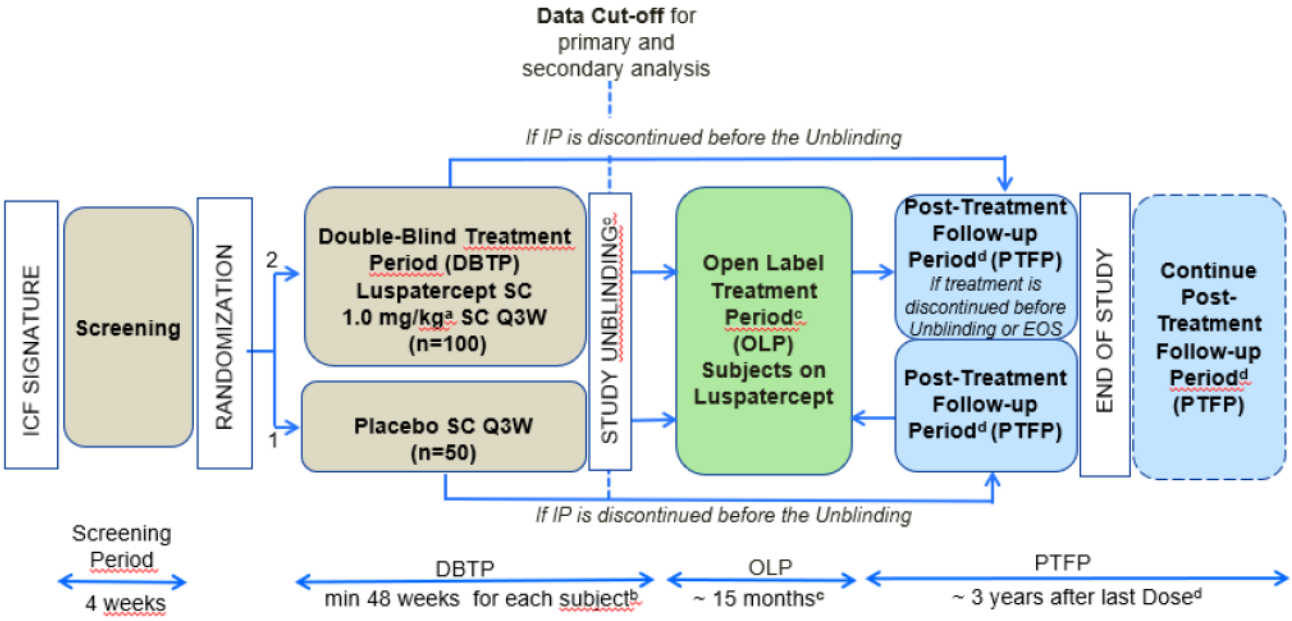
of IP, or 3 years from last dose (whichever occurs later), under the rollover protocol until the End of Trial. Specific assessments and visits to be performed during this period are defined in Table 3 of Protocol, Table of Events. After the unblinding, and after DMC's recommendation:

- Subjects who received placebo and have been assessed as per protocol at least up to 48 weeks after the first dose of IP (even if the IP is discontinued before completing 48 weeks of treatment), may stop the PTFP and access the OLP to receive luspatercept, if eligible.
- Subjects who received luspatercept and discontinued the IP in the DBTP or OLP may complete the PTFP and continue or complete the long-term follow-up of 5 years from first dose of IP, or (3 years after from last IP dose (whichever occurs later) to complete the PTFP under) in the rollover protocol until the End of Trial. Note: only the visit at 9 weeks after last dose should be performed in this study before moving to the rollover study.

The End of Trial is defined as the date all subjects complete the OLP (if allowed to access the OLP) , or discontinue earlier, or, the date of receipt of the last data point from the last subject that is required for primary, secondary, and/or exploratory analysis, as prespecified in the protocol, whichever is the later date.

The Sponsor may end the trial when all key endpoints and objectives of the study have been analyzed, and the availability of a rollover protocol exists into which any subjects remaining on study may be consented and continue to receive access to luspatercept, if not yet commercially available, and/or complete the PTFP.

Figure 1: Overall Study Design



EOS: end of study; EOT: end of trial; ICF: informed consent form; IP: investigational product; Q3W: every 3 weeks; SC: Subcutaneous.

[a] Dose may be titrated up to a maximum of 1.25 mg/kg.

[b] Double-blind Treatment Period (DBTP) will end after last subject enrolled has completed 48 weeks of treatment or discontinued earlier, or when the study is unblinded.

[c] The study will be unblinded 48 weeks after last subject has received the first dose of IP. At that time, subjects still benefitting from luspatercept treatment as well as subjects who received placebo and have been assessed as per protocol up to 48 weeks after the first dose of IP (even if they have discontinued the IP before completing 48 weeks of treatment), may access the OLP to receive luspatercept for maximum 15 months on the basis of DMC recommendation after unblinded data review, and can continue treatment in the rollover protocol after EOT up to 5 years of Dose 1, or treatment discontinuation, whichever occurs later.

[d] Subjects in the DBTP who have discontinued luspatercept before the unblinding or the OLP will continue the PTFP until the End of Trial (EOT) and may continue the PTFP in the rollover study up 5 years from first dose of IP, or 3 years from last dose (whichever occurs later), to complete the Post-treatment Follow-up Period under the rollover protocol. Subjects in the DBTP who have discontinued the placebo before the unblinding will continue the PTFP until the unblinding and may access the OLP, after DMC's recommendation.

4.2. Study Endpoints

4.2.1. Primary Efficacy Endpoint

The primary efficacy endpoint is:

- Proportion of subjects who have an increase from baseline ≥1.0 g/dL in mean of hemoglobin values over a continuous 12-week interval from Week 13 to Week 24 in the absence of transfusions.

4.2.2. Secondary Efficacy Endpoints

4.2.2.1. Key Secondary Efficacy Endpoints

The analyses of secondary efficacy endpoints will be performed on the ITT population. The key secondary endpoints will be measured from Dose 1 Day 1 of the DBTP, and will be statistically tested in a sequential order at two-sided $\alpha = 0.05$ level (details related to multiplicity adjustment can be found in Section 10.1). The key secondary efficacy endpoints include:

- Mean change from baseline in NTDT-PRO Tiredness and Weakness (T/W) domain score over a continuous 12-week interval from Week 13 to Week 24.
- Mean hemoglobin change from baseline over a continuous 12-week interval from Week 13 to 24 in the absence of transfusion.
- Proportion of subjects who have an increase from baseline ≥1.0 g/dL in mean of hemoglobin values over a continuous 12-week interval from Week 37 to Week 48 in the absence of transfusions.

4.2.2.2. Other Secondary Efficacy Endpoints

Other secondary efficacy endpoints include:

- Mean change from baseline in mean Functional Assessment of Chronic Illness Therapy-Fatigue (FACIT-F) Fatigue Subscale (FS) score over a continuous 12-week interval from Week 13 to Week 24.
- Mean change from baseline in mean NTDT-PRO Shortness of Breath (SoB) domain score over a continuous 12-week interval from Week 13 to Week 24
- Mean change from baseline in mean hemoglobin values in the absence of transfusions over a continuous 12-week interval from Week 37 to Week 48.
- Mean change from baseline in mean FACIT-F Fatigue Subscale (FS) score, mean NTDT-PRO T/W domain score and mean NTDT-PRO SoB domain score over a continuous 12-week interval from Week 37 to Week 48.

- Proportion of subjects with an increase from baseline ≥ 3 in mean FACIT Fatigue Subscale (FS) score over a continuous 12-week interval from Week 13 to Week 24.
- Proportion of subjects with an increase from baseline ≥ 3 in mean FACIT Fatigue Subscale (FS) score over a continuous 12-week interval from Week 37 to Week 48.
- Mean change from baseline in the physical component summary (PCS) and mental component summary (MCS) scores of Medical Outcomes Study 36-Item Short Form (SF-36-v2) at Week 24 and Week 48.
- Proportion of subjects with improvement of iron overload at Week 24 and Week 48, as measured by:
    - For subjects with baseline liver iron concentration (LIC) (by magnetic resonance imaging [MRI]) ≥3 mg/g dry weight (dw): ≥20% reduction in LIC, OR ≥ 33% decrease in iron chelation therapy (ICT) daily dose
    - For subjects with baseline LIC (by MRI) < 3 mg/g dw: no increase in LIC > 1 mg/g dw AND not starting treatment with ICT or no increase in ICT daily dose ≥33%, if on ICT at baseline.
- Mean change from baseline in serum ferritin at Week 24, Week 48, and up to last assessment.
- Mean change from baseline in LIC at Week 24, Week 48, and up to last assessment.
- Proportion of subjects who are transfusion-free over 24 weeks.
- Proportion of subjects who are transfusion-free over 48 weeks.
- Duration of the mean hemoglobin increase from baseline ≥1.0 g/dL.
- Mean change from baseline in the 6MWT distance at Week 24 and Week 48.
- Proportion of subjects who have an increase from baseline ≥1.5 g/dL in mean of hemoglobin values over a continuous 12-week interval from Week 13 to Week 24 in the absence of transfusions.
- Proportion of subjects with a decrease from baseline ≥ Responder Definition (RD) in mean NTDT-PRO T/W score, over Weeks 13 to 24 and Weeks 37 to 48

4.2.4. Safety Endpoints

Safety endpoints include:

- Type, frequency, and severity of adverse events and their relationship to investigational product (IP) (per NCI CTCAE version 4.0)
- Frequency of antidrug antibodies and its effect on efficacy and safety

4.3. Stratification, Randomization, and Blinding

Subjects will be randomized to receive luspatercept or placebo at a 2:1 ratio. Randomization will be accomplished by an IRT to ensure timely registration and randomization. A stratified block randomization schedule will be implemented. Randomization will be stratified by:

- Baseline hemoglobin level
  a. ≥ 8.5 g/dL
  b. < 8.5 g/dL
- Baseline NTDT-PRO T/W score
  a. ≥ 3 points
  b. < 3 points

4.4. Sample Size Determination

Based on the assumption of targeted primary endpoint response rates of at least 50% in the luspatercept group and 10% for the placebo group, and 2:1 randomization, a total sample size of

150 (100 in the luspatercept group, 50 in the placebo group) will have at least 99% power to detect the difference between the 2 groups with a 2-sided alpha of 0.05 and assumed 10% dropout rate.

For NTDT-PRO T/W domain scores, assume the mean change from baseline scores at Week 24 are 1.2 and 0.5 for luspatercept and placebo group, respectively, with a common standard deviation of 1.2, the statistical power will be 91%.

5. GENERAL STATISTICAL CONSIDERATIONS

5.1. Reporting Conventions

Summary tables, listings, figures and any supportive SAS outputs will include a "footer" of explanatory notes that will indicate, at a minimum, the following:

- Program source (e.g., SAS program name, including the path, that generates the output) and
- Data extraction date (e.g., the database lock date, run date)

The purpose of the data extraction date is to link the output to a final database, either active or archived, that is write-protected for replication and future reference. An output date will also appear on each output page and will indicate the date the output was generated by the analysis program.

The following reporting conventions apply generally to tables, listings, and figures:

- Data from all study centers will be combined for analysis;
- The randomization stratification variables used in the analyses will be derived based on the electronic case report form (eCRF) data.
- All statistical tests of the treatment effect will preserve a significance level of 0.050 for 2-sided tests. Testing of interactions will be performed at the 0.100 significance level, unless specified otherwise;
- P-values will be rounded to 4 decimal places. P-values that round to 0.0000 will be presented as '<0.0001' and p-values that round to 1.000 will be presented as '>0.9999';
- Confidence intervals (CIs) will be presented as 2-sided 95% CIs unless specified differently in specific analysis;
- Summary statistics will consist of the number and percentage of subjects in each treatment group for discrete variables, and the sample size, mean, median, Standard Deviation (SD), Q1, Q3, minimum, and maximum for continuous variables;
- All mean, median, Q1, and Q3 values will be formatted to one more decimal place than the measured value. Standard deviation values will be formatted to two more decimal places than the measured value; Minimum and maximum values will be presented to the same number of decimal places as the measured value.
- All percentages will be rounded to one decimal place. The number and percentage of responses will be presented in the form XX (XX.X%), where the percentage is in the parentheses; when the number of a response is zero, percentage will not be presented for that response;
- All listings will be sorted for presentation in order of treatment group, study center, subject, and date of procedure or event if not otherwise specified;

- All listings will display original collected values, cases with special marks (i.e., <500) will be listed as it is. The special mark will be removed if the value is used for calculation in tables;
- All analysis and summary tables will have the analysis population sample size (ie, number of subjects) if not otherwise specified;
- All summary tables will be displayed by treatment ("Luspatercept" and "Placebo"), the "Total" group will be added for sections if specified;
- In general, if not otherwise specified, baseline value will be defined as the last value (including "unscheduled") on or before the first dose of IP (if collecting time is available, date/time will be used to compare with first dosing date/time to identify baseline record, if there is no time available, only date will be used); if multiple values are present for the same date/time, the average of these values will be used as the baseline. For subjects who were not treated, the value on or prior to randomization date will be used. Specifically, for the laboratory hematology parameter 'Leukocytes', the baseline is defined as the highest value between screening visit and dose 1 day 1 visit.
- For data handling in change from baseline and shift tables (except for MRI and DXA parameters), "unscheduled" visits will be grouped with the closest scheduled visit based on assessment date. The average will be used as value for that scheduled visit in change from baseline tables; the worst category will be used in shift tables. If an unscheduled visit has equal distance to two scheduled visits, it will be grouped with the later visit. Specifically, for anti-drug antibody (ADA) titer summary, the titer value will not be averaged if an "unscheduled" visit is mapped to the closest scheduled visit. The titer for "unscheduled' visit will only be used for summary if the original scheduled visit has no titer result.
- For hemoglobin (Hb) concentration related efficacy endpoints summary, specific intervals are defined as:
  - Baseline 4-week interval: from Week -4 (Day -27 with an allowable administrative 3-day window) to Dose 1 Day 1 (Day 1). The details of the baseline definition will be specified in the programming specification.
  - Weeks 13 - 24 interval is chosen in the following order:
    a. Average of all Hb measurements in Weeks 13-24 (Day 86 to Day 169) if Weeks 13-24 has >=3 Hb measurements
    b. Average of all Hb measurements in Weeks 12-23 (Day 79 to Day 162) if Weeks 12-23 has >=3 Hb measurements
    c. Average of all Hb measurements in Weeks 14-25 (Day 93 to Day 176) if Weeks 14-25 has >=3 Hb measurements
    d. When the number of Hb measurements is less than 3 among Weeks 13-24, Weeks12-23, and Weeks14-25, then choose the largest number of Hb measurements in the following order as Hb at Week 13-24:
       1) Hb in Week 13-24
       2) Hb in Week 12-23

3) Hb in Week 14-25

e. Otherwise, missing

- Weeks 37 - 48 interval: Same rule as Week 13-24 interval

- The calculation of weekly NTDT-PRO scores is detailed in Section 18.3.1. The 12-week interval score is obtained by averaging all the weekly scores (at least one non-missing) in these 12 weeks. If all the weekly scores in these 12 weeks are missing, the 12-week interval score is then declared missing.

This SAP addresses efficacy and safety endpoints during the double-blinded treatment period (including the initial 48-week treatment period for all subjects and the long-term treatment period for early enrolled subjects). The database will be locked upon all subjects completed 48 weeks of the double-blind treatment period or discontinued before reaching 48 weeks. All the data collected up to data cutoff date will be used for summary. Data selection rules will be applied in each summary panel as needed, please refer to individual part for details.

5.2. Analysis Populations

A summary of analysis populations will be presented by treatment group and total.

5.2.1. Intent-to-Treat Population

The intent-to-treat (ITT) population will consist of all randomized subjects regardless of whether or not the subject received IP. All efficacy analyses will be conducted for the ITT population and will be analyzed based on randomization treatment.

5.2.2. Per Protocol Population

Subjects in the ITT who have taken at least 1 dose of IP and do not have important protocol deviations as described in Section 7 prior to database lock. The Per Protocol population might be used in addition to ITT population to analyze primary and key secondary endpoints.

5.2.3. Safety Population

The safety population will consist of all subjects who were randomized and received at least one dose of IP. Subjects will be included in the treatment group corresponding to the IP they actually received.

5.2.4. Health-related QoL Evaluable Population

The Health-related QoL (HRQoL) evaluable population consists of all subjects in the ITT population with a valid health-related QoL assessment at baseline (screening) and at least one valid postbaseline assessment.

6. SUBJECT DISPOSITION

The total number of subjects screened and total number of subjects with screen failure will be summarized and listed. Reasons subjects did not qualify for the study will be displayed by category. A corresponding listing will be provided.

A summary of subject disposition will be presented by treatment group and total for ITT, Per Protocol, HRQoL and safety populations.

Subject disposition summary will present the number and percentage of subjects for the following categories:

- subjects who were randomized,
- subjects who received treatment,
- subjects who discontinued study treatment,
- subjects whose treatment were ongoing,
- subjects who completed 24 weeks of treatment,
- subjects who completed 48 weeks of treatment,
- subjects who discontinued from the study by treatment group and total.
- the reasons for discontinuation of study treatment
- the reasons for discontinuation of study participation

All percentages will be based on the number of subjects randomized using the ITT, Per Protocol, HRQoL, or safety populations.

The reasons for treatment discontinuation will be collected on the electronic case report form (eCRF) and summarized for all treated subjects based on the following categories:

- Death
- Adverse event
- Pregnancy
- Progressive disease
- Lack of efficacy
- Recovery
- Withdrawal by subject
- Non-compliance with study drug
- Lost to follow-up
- Study terminated by sponsor

- Transition to commercially available treatment
- Physician decision
- Site terminated by sponsor
- Protocol deviation
- Transition to rollover Protocol
- Failure to meet treatment criteria
- Other: dose is delayed for more than 15 weeks due to AEs (including case of elective surgery/hospitalization
- Other

The reasons for study discontinuation will be collected on the eCRF (only when FUP period is not completed) and will be summarized for all randomized subjects based on the following categories:

- Death
- Adverse event
- Pregnancy
- Lack of efficacy
- Withdrawal by subject or parent/guardian
- Non-compliance with study drug
- Lost to follow-up
- Study terminated by sponsor
- Transition to commercially available treatment
- Physician decision
- Disease relapse
- Protocol deviation
- Transition to rollover protocol
- Other

A summary of subjects enrolled by geographic region, country, and site will be provided in a separate table by treatment group and total.

A subject disposition listing will be provided.

7. PROTOCOL DEVIATIONS AND IMPORTANT PROTOCOL DEVIATIONS

The protocol deviations and important protocol deviations will be identified and assessed by clinical research physician or designee following company standard operational procedure. An important protocol deviation occurs when there is any departure from the approved protocol that: impacts the safety, rights, and/or welfare of the subject; or negatively impacts the quality or completeness of the data; or makes the informed consent document/form inaccurate. Important protocol deviations are identified based on blinded data reviews of deviation log throughout the study and are finalized prior to database lock.

The number and percentage of the subjects with any protocol deviation, or important protocol deviation will be provided for the ITT population respectively by treatment group and total. For important protocol deviations, the number and percentage of subjects within each subcategory will be summarized as well.

All protocol deviations will be listed for the ITT population.

8. DEMOGRAPHICS AND BASELINE CHARACTERISTICS

The demographics and baseline characteristics will be summarized for the ITT population. Individual subject listings will be provided to support the summary tables.

8.1. Demographics

Summary statistics will be provided descriptively by treatment group and total for the following continuous variables:

- Age
- Weight (kg)
- Height (cm)
- Body mass index (BMI; kg/m^2)

Age or date of birth will be recorded on eCRF. Where age is not recorded, age will be calculated as described in Section 18.1.1.

Body mass index will be calculated as follows: BMI (kg/m^2) = baseline weight in kg / (height in $m)^2$.

A frequency summary (number and percentage) will be provided by treatment group for the following categorical variables:

- Age category ($\leq$32 years, > 32 - $\leq$ 50 years and > 50 years)
- Sex (Male, Female with or without childbearing potential, Unknown, Undifferentiated)
- Race (American Indian or Alaska Native, Asian, Black or African American, Native Hawaiian or Other Pacific Islanders, White, Not Collected or Reported, Other)
- Ethnicity (Hispanic or Latino, Not Hispanic or Latino, Not Reported, Unknown)
- Region (North America and Europe, Middle East, Asia-Pacific)
- BMI category (< 20, $\geq$ 20 to < 25, $\geq$ 25 to < 30, $\geq$ 30 kg/m^2)

8.2. Baseline Characteristics

The following baseline characteristics will be summarized.

- Beta-thalassemia diagnosis: Beta-Thalassemia, Hemoglobin E/Beta-Thalassemia, Beta-Thalassemia combined with Alpha-Thalassemia;
- Baseline hemoglobin level (mean of at least 2 Hb values by central lab at least 1 week apart during the 28 days screening period) summarized as continuous variable as well as by category ($\geq$ 8.5 g/dL and < 8.5 g/dL) excluding the Hb values within 21 days following a transfusion;

- Baseline NTDT-PRO T/W Score (calculated over 7 days prior to Dose 1 Day 1 detailed in Section 18.3.1) summarized as continuous variable as well as by category (≥ 3 points and < 3 points);
- Baseline NTDT-PRO SoB Score (calculated over 7 days prior to Dose 1 Day 1 detailed in Section 18.3.1);
- Baseline transfusion burden in units/24 weeks before Dose 1 Day 1 (descriptive and categorized level: =< median value and > median value if baseline transfusion burden > 0);
- Baseline transfusion burden in units/12 weeks before Dose 1 Day 1 (descriptive and categorized level: =< median value and > median value if baseline transfusion burden > 0);
- Beta-thalassemia gene mutation grouping: Beta Allele 1 Mutation, Beta Allele 2 Mutation, Alpha Gene 1 Mutation, Alpha Gene 2 Mutation, Other Disease-related Mutations
- Eastern Cooperative Oncology Group (ECOG) performance status (0 or 1) at screening visit to assess how the disease affects subjects' daily activities: 0 = fully active, able to carry on all pre-disease performance without restriction; 1 = restricted in physically strenuous activity but ambulatory and able to carry out work of a light or sedentary nature, e.g., light housework, office work; A listing will be provided for ECOG data.
- Splenectomy: Yes/No;
- Baseline Serum Ferritin;
- Hepatitis B and C results;
- MRI liver iron content (LIC) (descriptive and categorized level: <3, >= 3-≤ 5, > 5-≤ 7, >7-≤ 15 and > 15). The value of LIC will be either the value collected from eCRF or the value derived from T2*, R2* or R2 parameter depending on which techniques and software was used for MRI LIC acquisition. Please refer to Section 10.4.10 for more imputation details
- Bone mineral density DXA scan (BMDs and T-scores by location);
- Baseline 6-Minute Walk Test (6MWT);
- Baseline FACIT-F Fatigue Subscale (FS);
- Baseline SF-36 PCS and MCS Scores.

8.3. Beta-thalassemia Comorbidities/Medical History

The Beta-thalassemia comorbidities and medical history will be coded by Medical Dictionary for Regulatory Activities (MedDRA; Version 23.0), and summarized by system organ class (SOC) and preferred term (PT) by treatment group and total. Separate tables will be provided for current

Beta-thalassemia comorbidities and the history of Beta-thalassemia comorbidities. The SOCs and PTs will be listed in descending frequency within the luspatercept group. A subject will be counted only once for multiple events within each SOC/PT.

A separate table will be provided to summarize Beta-thalassemia comorbidities by comorbidity terms for each treatment group and total.

Corresponding listing will be provided. For cancer history, data will be presented in the listing.

[redacted]

9 STUDY TREATMENTS AND EXTENT OF EXPOSURE

Subjects will be assigned to one of following regimens during treatment phase:

- Luspatercept starting dose level 1.0 mg/kg SC once every 21 days
- Placebo SC once every 21 days

Study treatment and extent of exposure summaries will be provided based on the safety population. Descriptive statistics will be provided for treatment duration, number of doses received per subject/treatment and average number of days per dose by treatment group and total. The number and percentage of subjects will be summarized for maximum dose level received, and dose level reduced to by treatment group and total. Corresponding listing will be provided.

9.1 Treatment Duration

Treatment duration (weeks) is defined as:

[(The treatment end date) – (The treatment start date) + 1]/7,

where the treatment start date is the date of the first dose of study drug. The treatment end date is min [(date of last dose + 20), death date]. For subjects who are still on treatment at the time of study closure or clinical cutoff, the minimum of (data cutoff date, date of last dose+ 20) will be used as the treatment end date.

Descriptive statistics will be summarized for treatment duration by treatment group and total.

9.2 Number of Doses Received per Subject

Total number of doses received per subject is defined as the total number of doses the subject received (i.e., total number of non-zero doses). It will be summarized descriptively and categorically (1, 2, 3, 4, 5, 6, 7, 8, 9 – 16, 17 – 24, 25 – 32, >32) by treatment group and total. The total number of doses received will be calculated from all subjects within each treatment group. The total number and percentage of doses received for each planned dose level (1.25 mg/kg, 1.0 mg/kg, 0.80 mg/kg, 0.60 mg/kg, 0.45 mg/kg) will be calculated within treatment group, with the total number of doses received as denominator.

9.3 Average Number of Days Between Doses

Average number of days between doses is defined as number of days on treatment (treatment duration) divided by the number of doses (where a subject received a non-zero dose) for each subject. Descriptive summary statistics will be provided for average number of days per dose by treatment group and total.

9.4 Dose Modifications: Dose Delay, Dose Titration and Dose Reduction

9.4.1 Dose Delay

Does delay is defined as delay of planned dose schedule due to increased hemoglobin (≥ 11.5 g/dL) or adverse events which also further categorized by related events equal to grade 2 and ≥ grade 3 or WBC count ≥ 3X baseline or any other reasons. If dose delay exceeds 15 weeks from last dosing date, the treatment should be discontinued.

9.4.2 Dose Titration and Dose Reduction

Dose adjustment includes dose reduction and dose titration (increase). Titration is based on hemoglobin level over the previous two doses compared to baseline. Subjects may be dose-escalated up to 1.25 mg/kg during the DBTP and OLP, but the maximum total dose per administration should not exceed 120 mg. The dose escalation criteria are defined as follows:

- Dose escalation may be performed if at a constant dose level, the increase of mean hemoglobin (uninfluenced by transfusions, ie, > 21 days post-transfusion) over 2 cycles (6 weeks) is < 1.0 g/dL, compared to the baseline hemoglobin value (mean of 2 values 1 week apart within 4 weeks of randomization);
- Dose escalation may be performed if at a constant dose level, the increase of mean hemoglobin (uninfluenced by transfusions, ie, > 21 days post-transfusion) over 2 cycles (6 weeks) is ≥ 1.0, but < 2.0 g/dL compared to the baseline hemoglobin value (mean of 2 values 1 week apart within 4 weeks of randomization).

Dose reduction can be caused by increased or high level of hemoglobin or adverse events. Subjects who have been dose-reduced due to any related AE ≥ Grade 3, as indicated in Table 5 of Protocol, should not be dose-escalated during the DBTP. Although, following Investigator's requests, the Sponsor may allow dose increase to the next higher dose level, after safety and efficacy data review. Starting dose with dose reductions and escalations are presented in Table 1.

Table 1: Starting Dose Level with Dose Adjustments

3rd Dose Reduction (~ 25 %)	2nd Dose Reduction (~ 25 %)	1st Dose Reduction (~ 25 %)	Starting Dose Level	1st Dose Increase
0.45 mg/kg	0.6 mg/kg	0.8 mg/kg	1.0 mg/kg	1.25 mg/kg

The dose delay, dose reduction, and dose titration will be summarized in separate tables by treatment group and total. The number of subjects with at least one dose delay/reduction/titration, number of dose delays/reduction/titration per subject, reason for dose delay and adjustment, time to first dose delay/reduction/titration (days), and time to first dose delay/reduction due to AE (days) will be summarized by treatment group and total. Corresponding listing will be provided.

9.5 Investigational Drug Overdose

Overdose (refers to luspatercept only), is defined as SC 10% over the protocol-specified dose level assigned to a given subject, regardless of adverse events or sequelae. A listing will be provided for any overdose, which occurs accidentally or intentionally as collected in the eCRF.

10 EFFICACY ANALYSIS

All efficacy evaluations will be conducted using the ITT population with the exception of FACIT-F Fatigue Subscale (SF) and SF-36 analyses that will be conducted on the HRQoL evaluable population. NTDT-PRO T/W and SoB scores will be analyzed by both ITT and HRQoL populations. The primary and key secondary endpoints will also be analyzed based on the Per Protocol population as a sensitivity analysis.

Some continuous variables like hemoglobin and weekly NTDT-PRO T/W and SoB domains will be derived as the sum of results over 12 weeks divided by the number of assessments. The mean over 12 weeks will be used to derive mean change from baseline for each subject.

For the early treatment-discontinued subjects, i.e. subjects who did not complete 24 weeks or 48 weeks of double-blinded treatment period, the transfusion and assessment records (including Hb and QoL) will still be collected up to 48 weeks post Dose 1 Day 1 or 9 weeks post last dose, whichever is the later date. All the assessments before the efficacy cutoff date will be used to evaluate primary and secondary endpoints. For these endpoints, if at the time of data summary, a subject has died or has not reached the end of the 12-week interval, this subject will be included in the analysis. For other efficacy endpoints, the collected records will be used (by visits or at specified time point) as needed, depending on the purpose of the summary. Please refer to the individual endpoint section for details.

The efficacy cutoff date is defined as the minimum date among death date, study discontinuation date, last dose date + 20, and data cutoff date. For the primary and responder-related secondary endpoints, if at the time of data summary, a subject's efficacy cutoff date is before the end of the 12-week interval or a subject has any invalid hemoglobin records (i.e., within 21 days following a transfusion) during the specified 12-week interval, this subject will be included in the analysis and his/her responder status will be classified in accordance with the definition of erythroid response in Section 10.2. For other efficacy endpoints, if a subject's efficacy cutoff date is before the end of the 12-week interval, the subject will be included in the analyses.

Statistical comparisons will be made between luspatercept and placebo. Key secondary efficacy results that will be considered statistically significant after consideration of the strategy for controlling the family-wise Type 1 error rate are described in Section 10.1. All statistical tests will be 2-sided at the significance level of $\alpha = 0.05$, and the corresponding p-values and 2-sided CIs for point estimates will be reported, unless specified otherwise.

In general, descriptive statistics will be provided and statistical tests will be applied if appropriate. For continuous variables, least squares (LS) means with corresponding standard errors (SE) for each treatment group, along with LS mean of treatment difference (luspatercept versus placebo) with corresponding 95% CI and p-value will be presented for analysis of covariance (ANCOVA). Counts and percentages will be used to describe categorical variables, and the treatment comparison will be analyzed analogous to the primary efficacy endpoint, using the CMH model

stratified by baseline randomization stratification factor(s) The odds ratio (OR) (luspatercept versus placebo) with corresponding 2-sided (at 0.05 alpha level) 95% CI and p-value will be provided. The difference in proportions (luspatercept – placebo) and 95% CI will also be calculated. A forest plot showing the ORs, 95% CI and p-value for the overall result and the results in each subgroup will be constructed for primary and key secondary endpoints. For all endpoints mentioned in this section, records beyond week 48 will not be used unless otherwise specified. If ANCOVA is used, the statistical assumption will be validated first, log transformation will be applied as needed.

10.1 Multiplicity

Gate-keeping methods will be used to control the overall Type 1 error rate for the key secondary endpoints. After the result from the primary efficacy analysis in the ITT population shows statistical significance, the key secondary endpoint 1 will be tested next. The key secondary endpoint 2 will be tested only if the test results for both primary endpoint and the key secondary endpoint 1 are significant. The key secondary endpoint 3 will be tested only if the test results for primary endpoint and the key secondary endpoints 1 and 2 are all significant. The testing procedure above will be implemented strictly in order to control the overall Type 1 error rate of 0.05 due to multiplicity.

10.2 Analysis of Primary Efficacy Endpoint

The primary efficacy endpoint of this study is erythroid response, defined as an increase from baseline ≥1.0 g/dL in mean of hemoglobin values over a continuous 12-week interval from Weeks 13 to 24 in the absence of transfusions. The response rate is defined as the number of responders divided by the number of subjects in the ITT population within each treatment group.

Baseline hemoglobin (Hb) is the average of 2 or more Hb measurements at least 1 week apart within 4 weeks prior to Dose 1 Day 1. The Hb values within 21 days following a transfusion may be influenced by the transfusion and will be excluded from this analysis. For discontinued subjects who do not complete 24 weeks of the DBTP, Hb data will continue to be collected. For a subject, if the Hb average over Weeks 13-24 interval cannot be calculated due to missing data, the closest Hb 12-week average will be used as mentioned in Section 5.1. If the 12-week average is still missing, the imputation method will be performed as described in Section 10.6.1 and the responder status will be classified in accordance with the definition of erythroid response in Section 10.2.

The number and percentage of subjects in the ITT population who achieve the response will be calculated for luspatercept and placebo. The primary efficacy analysis for treatment comparison (luspatercept versus placebo) will be conducted by the Cochran Mantel-Haenszel (CMH) test with the treatment and randomization stratification factor(s) in the model. The odds ratio (OR) (luspatercept versus placebo) with corresponding 95% confidence interval (CI) and p-value will be calculated with treatment group, randomization stratification factor(s) in the model to compare the luspatercept and placebo groups at 2-sided 0.05 level. In addition, the difference in response rate

between luspatercept and placebo with corresponding 95% CI will be calculated by exact unconditional test.

A forest plot showing the ORs, 95% CI and p-value for the overall result and the results in each subgroup will be constructed. Listing of individual Hb data will be provided.

The subgroup analyses described in Section 10.5. including mean baseline Hb (>=8.5 g/dL and <8.5 g/dL) for the primary endpoint will also be performed by summarizing the response rate difference and OR with corresponding 95% CI.

A sensitivity analysis will be conducted by classifying subjects with less than 3 Hb measurements from Week 13 (-7 days) to Week 24 (+7 days) as non-evaluable and excluded in the analysis. Additional sensitivity analysis will be conducted by classifying those non-evaluable subjects as non-responders. The same sensitivity analysis will be followed for Hb secondary endpoints over Weeks 37 to 48. Additional sensitivity analyses for the missing assessments are stated in Section 10.6.1.

10.3 Analyses of Key Secondary Efficacy Endpoints

The analyses of key secondary efficacy endpoints will be performed on the ITT population. The key secondary endpoints will be tested in the order as mentioned from Section 10.1 and listed below:

1 Mean change from baseline in NTDT-PRO Tiredness and Weakness (T/W) domain score over a continuous 12-week interval from Week 13 to Week 24.

2 Mean hemoglobin change from baseline over 12-week interval from Week 13 to 24 during the treatment period in the absence of transfusions.

3 Proportion of subjects who have an increase from baseline ≥1.0 g/dL in mean of hemoglobin values over a continuous 12-week interval from Week 37 to Week 48 in the absence of transfusions.

To control the overall Type 1 error rate for secondary endpoints, the testing procedure will be implemented strictly in order: the test for key secondary endpoint 1 will only be conducted when there is evidence showing that erythroid response is achieved in the luspatercept group from Week 13 to Week 24 (primary endpoint). The key secondary endpoint 2 will be tested only if the test results for both primary endpoint and the key secondary endpoint 1 are significant. The key secondary endpoint 3 will be tested only if the test results for primary endpoint and the key secondary endpoints 1 and 2 are all significant.

10.3.1 NTDT-PRO (T/W Domain) Mean Change Between Weeks 13 to 24

The change from baseline on mean of NTDT-PRO (T/W domain) over a 12-week interval between Weeks 13 to 24 will be analyzed using analysis of covariance (ANCOVA) method with treatment group in the model and randomization stratification factor(s) as covariates. Treatment effect will be evaluated as a contrast of luspatercept versus placebo. Least squares (LS) means with

corresponding standard errors (SE) for each treatment group, along with LS mean of treatment difference with corresponding 95% CI and p-value will be presented.

The subgroup analyses including baseline NTDT-PRO T/W (≥ 3 and <3) will be performed as described in Section 10.5.

The treatment-by-stratification-factor(s) interaction will be added to the ANCOVA model as a sensitivity analysis and results from this model will be compared to the primary fixed effects model (without the interaction effect). If the interaction is significant at a 2-sided 0.1-alpha level, the nature of this interaction will be inspected as to whether it is quantitative (that is, the treatment effect is consistent in direction but not in size of effect) or qualitative (the treatment is beneficial in some values of the stratification factor(s)). If the interaction effect is found to be quantitative, results from the primary fixed effect model will be presented. If the interaction effect is found to be qualitative, further inspection will be used to identify in what values of the stratification factor(s) the treatment is more beneficial.

10.3.2 Hemoglobin Mean Change Between Weeks 13 to 24

Mean Hb change from baseline over a 12-week interval between Weeks 13 to 24 will be analyzed using ANCOVA method with treatment group in the model and randomization stratification factor(s) as covariates. Treatment effect will be evaluated as a contrast of luspatercept versus placebo. Least squares (LS) means with corresponding standard errors (SE) for each treatment group, along with LS mean of treatment difference with corresponding 95% CI and p-value will be presented.

Baseline hemoglobin (Hb) is the average of 2 or more Hb measurements at least 1 week apart within 4 weeks prior to Dose 1 Day 1. The Hb values within 21 days following a transfusion may be influenced by the transfusion and will be excluded from this analysis. For discontinued subjects who do not complete 24 weeks of the DBTP, Hb data will continue to be collected. For a subject, if the Hb average over Weeks 13-24 interval cannot be calculated due to missing data, the closest Hb 12-week average will be used as described in Section 5.1. If the value is still missing, the imputation method will be performed as described in Section 10.6.1.

A waterfall plot will be provided for the hemoglobin change from baseline by treatment group. Individual subject's hemoglobin change from baseline will be displayed in a single bar. The displayed hemoglobin change is each subject's largest decrease in hemoglobin.

The subgroup analyses including mean baseline Hb (>=8.5 g/dL and <8.5 g/dL) will be performed, as described in Section 10.5.

10.3.3 Hemoglobin Response Between Weeks 37 to 48

Hemoglobin responder, defined as an increase from baseline ≥ 1.0 g/dL in mean of hemoglobin values over a continuous 12-week interval from Weeks 37 to 48 in the absence of transfusions will be analyzed.

Baseline hemoglobin (Hb) is the average of 2 or more Hb measurements at least 1 week apart within 4 weeks prior to Dose 1 Day 1. The Hb values within 21 days following a transfusion may be influenced by the transfusion and will be excluded from this analysis. For discontinued subjects who do not complete 24 weeks of the DBTP, Hb data will continue to be collected. For a subject,

if the Hb average over Weeks 37-48 interval cannot be calculated due to missing data, the closest Hb 12-week average will be used.

The number and percentage of subjects in the ITT population who achieve the response will be calculated for luspatercept and placebo. The odds ratio (OR) (luspatercept versus placebo) with corresponding 95% CI and p-value will be provided and the CMH chi-square test will be conducted with treatment group and randomization stratification factor(s) in the model to compare the luspatercept and placebo groups at 2-sided 0.05 level. The difference in proportions (Luspatercept minus placebo) with corresponding 95% CI will be calculated by exact unconditional test.

The subgroup analyses including mean baseline Hb (>=8.5 g/dL and <8.5 g/dL) will be performed, as described in Section 10.5.

10.4 Analyses of Other Secondary Efficacy Endpoints

The analyses of other secondary efficacy endpoints will be performed on the ITT population except for the FACIT-F and SF-36 endpoints listed in this section. The analyses of NTDT-PRO, FACIT-F, and SF-36 endpoints will be conducted based on the Health Related QoL evaluable population (Section 5.2.4).

10.4.1 FACIT-F Fatigue Subscale (FS): Mean Change from Baseline between Weeks 13 to 24 and Weeks 37 to 48

Mean FACIT-F FS score change from baseline over a 12-week interval between Weeks 13 to 24 (nominal Dose 5 to Dose 9) will be analyzed using ANCOVA method with treatment group, randomization stratification factor(s) in the model and baseline FACIT-F FS score as covariate. The baseline of FACIT-F FS is the score from the last visit prior to or on Dose 1 Day 1.

Mean change from baseline for Weeks 37 to 48 (nominal Dose 13 to Dose 17) will be analyzed similarly.

The subgroup analyses for Weeks 13-24 will be performed as described in Section 10.5.

10.4.2 NTDT-PRO SoB Score: Mean Change from Baseline between Weeks 13 to 24 and Weeks 37 to 48

Mean NTDT-PRO SoB score change from baseline over a 12-week interval between Weeks 13 to 24 will be analyzed using ANCOVA method with treatment group, randomization stratification factor(s) in the model and baseline NTDT-PRO SoB score as covariate. Similar analysis will be performed for Weeks 37 to 48.

10.4.3 Hemoglobin Mean Change Between Weeks 37 to 48

Mean Hemoglobin change from baseline over 12-week interval between Weeks 37 to 48 during the treatment period will be analyzed using ANCOVA method with treatment group and randomization stratification factor(s) in the model as covariates.

Baseline hemoglobin (Hb) is the average of 2 or more Hb measurements at least 1 week apart within 4 weeks prior to Dose 1 Day 1. The Hb values within 21 days following a transfusion may be influenced by the transfusion and will be excluded from this analysis. For discontinued subjects who do not complete 24 weeks of the DBTP, Hb data will continue to be collected. For a subject, if the Hb average over Weeks 37-24 interval cannot be calculated due to missing data, the closest Hb 12-week average will be used.

10.4.4 NTDT-PRO (T/W Domain) Mean Change Between Weeks 37 to 48

Mean NTDT-PRO (T/W domain) change from baseline over a 12-week interval between Weeks 37 to 48 (nominal Dose 13 to Dose 17) will be analyzed using ANCOVA method with treatment group and randomization stratification factor(s) in the model as covariates.

The treatment-by-stratification-factor(s) interaction will be added to the ANCOVA model as a sensitivity analysis and results from this model will be compared to the primary fixed effects model (without the interaction effect). If the interaction is significant at a 2-sided 0.1-alpha level, the nature of this interaction will be inspected as to whether it is quantitative (that is, the treatment effect is consistent in direction but not in size of effect) or qualitative (the treatment is beneficial in some values of the stratification factor(s)). If the interaction effect is found to be quantitative, results from the primary fixed effect model will be presented. If the interaction effect is found to be qualitative, further inspection will be used to identify in what values of the stratification factor(s) the treatment is more beneficial.

10.4.5 FACIT-F Fatigue Subscale (FS) Response at Weeks 13 to 24 and Weeks 37 to 48

Proportion of subjects with an increase from baseline ≥ 3 in mean FACIT-F FS score, over Weeks 13 to 24 in the luspatercept group will be compared with the placebo group. The CMH test will be used with baseline FACIT-F FS category ≥ 37 versus < 37 as the stratum to compare the response rates between the 2 groups. The selection of 37 as a threshold is based on a Phase 2 Luspatercept study (A536-06) in which it corresponds to a NTDT-PRO T/W domain scores of 4. The corresponding 95% confidence interval for odds ratio will also be provided. Similar analysis will be performed for Weeks 37 to 48.

10.4.6 SF-36 v2: Mean Change from Baseline at Week 24 and Week 48

The SF-36 Version 2.0 (acute condition with 1-week recall period) is a self-administered instrument consisting of 8 multi-item scales that assess 8 health domains: Physical functioning (PF), Role-Physical (RP), Bodily Pain (BP), General Health (GH), Vitality (VT), Social functioning (SF), Role-Emotional (RE), and Mental Health (MH). Two overall summary scores, Physical Component Summary (PCS) and Mental Component Summary (MCS), will be calculated from the 8 health domains. The primary interests of the SF-36 are the PCS and MCS norm-based scores.

Scores for each domain and composite summary scale of the SF-36 will be derived using the Quality Metric's Health Outcomes™ Scoring Software 5.1, a software designed to provide standardized scoring methods for the SF-36 based on the 2009 US general population normative data (*Maruish, 2011*). The scoring algorithm, using the 1998 US general population normative data as an example, for the SF-36 is shown in Appendix 18.3.3.

Summary statistics (n, mean, median, SD, range) for PCS and MCS norm-based scores along with the change from baseline will be summarized at Weeks 12, 24, 36, and 48 for each treatment group. The Norm-base scores will be used for domain analysis and the PCS analysis. The raw scores for each individual question and the calculated domain and component scores will be presented in the listings for SF-36.

A mixed effect repeated measure model will be used with intercept and time as the random effect. The model includes the change from baseline scores up to Week 48 as the outcome, and treatment group assignment, time (weeks since Dose 1 Day 1), baseline scores, randomization stratification factor(s), time-by-baseline score, and time-by-treatment group assignment terms in the model. The unstructured covariance matrix will be used to obtain the random effects variance components.

If the MMRM model using an unstructured covariance matrix fails to converge with the default Newton-Raphson algorithm, the Fisher scoring algorithm or another appropriate method will be used to provide initial values of the covariance parameters. In the rare event that none of the above methods yields convergence, a structured covariance, such as heterogeneous Toeplitz and Toeplitz structures, will be used to model the correlation among repeated measurements. In this case, the empirical option will be used because the sandwich variance-covariance estimator is asymptotically consistent.

To assess the extent of missing data at each assessment visit by treatment group, compliance rates for the SF-36 will be estimated on the ITT population separately based on the number of subjects included in the ITT population per treatment group who are eligible for assessment at a given scheduled visit. Subjects will be considered compliant with completion of the SF-36 assessment if at least half of the 36 items (i.e., $\geq$ 18 items) are completed.

Additional details of the highlighted analyses as well as sensitivity analyses, and additional analyses to assess subgroups and treatment effects on HRQoL are provided in a separate HRQoL statistical analysis plan, which will be finalized prior to database lock. The HRQoL statistical analysis plan will be appended to the separate specific HRQoL report.

10.4.7 Mean Change in mean Daily Dose of Iron Chelation Therapy

The ICT mean daily dose summary will be provided for subjects who did not change ICT drug from baseline to post-baseline and only one ICT drug has been used. Descriptive statistics for mean daily dose will be summarized at baseline and the post-baseline visit for each ICT drug. The baseline mean daily dose will be calculated using the ICT dosage during the 24 weeks on or prior to first study drug treatment and the post-baseline mean daily dose is calculated during the last 24 weeks of the 48-week double-blind treatment period or the last 24weeks of the study treatment for early discontinued subjects.

The same descriptive statistic summary for baseline and post-baseline ICT drug mean daily dose will be provided for subjects in each baseline liver iron content category (<3; >=3-<=5; >5-<=7; >7-<=15; >15 mg/g dry weight).

The change from baseline in mean daily dose at post-baseline will be analyzed using an ANCOVA model with treatment group, randomization stratification factor(s) and baseline ICT mean daily dose as covariates based on the ITT population for subjects who did not change ICT drug from baseline to post-baseline and only one ICT drug has been used.

A summary showing the number and percentage of subjects who took monotherapy (i.e., only one ICT drug) vs. combo therapy (i.e., more than one ICT drug) at the 24-week baseline period and post-baseline period will be provided.

Bar plot will be provided for percent of subjects who took each ICT drug at baseline and postbaseline for subjects who did not change ICT drug from baseline to post-baseline and only one ICT drug has been used. A similar bar plot will be provided for percent of subjects who took monotherapy or combo therapy at baseline and post-baseline.

10.4.8 LIC / ICT: Responder Analysis at Week 24 and Week 48

There are 3 scheduled LIC measurements for each subject: baseline, Weeks 24 and 48. For subjects who discontinued before 48 weeks, the LIC and ICT data may be collected in the PTFP. The value of LIC will be either the value collected from eCRF or the value derived from T2* (see section 10.4.10 for further details). ICT daily dosing data will be collected at every visit during the study. Baseline ICT dose is based on medical history over 24 weeks prior to Dose 1 Day 1.

At Week 24, a LIC/ICT responder will be defined as one of the following:

- Baseline LIC (by MRI) ≥ 3 mg/g dw: ≥ 20% reduction in LIC, OR, ≥ 33% decrease in ICT daily dose
- Baseline LIC < 3 mg/g dw: no increase in LIC > 1 mg/g dw AND not starting treatment with ICT or no increase in ICT daily dose ≥ 33%, if on ICT at baseline

CMH test will be performed with treatment group, baseline ICT use (Yes, No) and randomization stratification factors as strata to compare the luspatercept and placebo groups at 2-sided 0.05 level; the corresponding 95% confidence interval for odds ratio will also be calculated.

The subgroup analyses will be performed as described in Section 10.5.

10.4.9 Serum Ferritin: Mean Change from Baseline at Week 24 and Week 48

Descriptive statistics for serum ferritin level will be summarized at baseline, at Week 24 and Week 48 of the 48-week DBTP. The baseline mean serum ferritin is measured during the 24 weeks prior to the first dose and the post-baseline mean serum ferritin is calculated during the last 24 weeks of the 48-week double-blind treatment period or the last 24 weeks of the study treatment for early discontinued subjects. Change from baseline will be summarized at Week 24 and Week 48.

The change from baseline in serum ferritin at Week 24 and Week 48 will be analyzed using an ANCOVA method with treatment group and randomization stratification factor(s) in the model and baseline Serum Ferritin as the covariate. If the normality assumption is significantly deviated, the nonparametric Wilcoxon rank-sum test will be used instead. Mean change at Week 48 will be analyzed similarly.

All the serum ferritin data will be presented in a listing.

10.4.10 Liver Iron Concentration: Mean Change from Baseline at Weeks 24 and 48

LIC mean change from baseline at Week 24 will be summarized descriptively and analyzed using ANCOVA method with treatment group and randomization stratification factor(s) in the model and baseline LIC as covariates.

LIC mean change from Baseline at Week 48 will be analyzed similarly.

The value of LIC will be either the value collected from eCRF or the value derived from T2*, R2* or R2 parameter depending on which techniques and software were used for MRI LIC acquisition and post-processing. The LIC value will be derived as below:

Technique	Site number	Derivation source	Alternative derivation if still missing
T2*/R2*	[redacted]	0.0254 * (R2*) + 0.2	25.4/T2* +0.2 If both R2* and T2* are missing, LIC from eCRF will be used
T2*/R2*	[redacted]	25.4/T2* + 0.2	If both T2* and R2* are missing, LIC from eCRF will be used
T2*/R2*	[redacted]	31.94 * (T2*)^-1.014	0.029*(R2*)^1.014 If both T2* and R2* are missing, LIC from eCRF will be used
Ferriscan R2	[redacted]	Reported LIC from CRF	(29.75-SQRT(900.7-2.283*EXP ((-0.19043+1.016385* ln (R2)) /0.983615))) ^1.4265

The derived LIC value will be used for analysis. If a subject has any LIC value > 43, the subject's LIC value will be excluded from analysis. Note that, subjects with LIC value > 43 are not excluded from the LIC baseline summary and efficacy subgroup analysis by LIC categories.

Descriptive statistics for LIC measurements and change from baseline will be summarized at week 24/48. The 24/48-week LIC change from baseline will be analyzed using ANCOVA model and randomization stratification factor(s) in the model and baseline LIC as covariates. Additionally, a shift table representing the shift from the baseline to week 24/48 category (<3; >=3-<=5; >5-<=7; >7-<=15; >15 mg/g dry weight) will be provided for LIC. A subject will have maximum two post-baseline LIC assessments (including "unscheduled") during the 48 weeks double-blinded treatment period per protocol. The assessment will be counted as the closest visit ("Week 24" or "Week 48") based on its actual assessment date. This logic will be used in model based summary, change from baseline summary and the shift table summary.

Additionally, bar plot will be provided for baseline, week 24 and week 48 LIC categories (<3; >=3-<=5; >5-<=7; >7-<=15; >15 mg/g dry weight). Percent of subjects within each category will be displayed by treatment group. All the LIC data will be presented in a listing.

10.4.11 Transfusion Free for 24 and 48 weeks

Proportion of subjects who receive no RBC transfusions from Week 1 to Week 24 will be summarized and compared between the 2 treatment groups using CMH test with randomization stratification factor(s) adjusted.

Similarly, proportion of subjects who receive no RBC transfusion from Week 1 to Week 48 will be compared between the 2 treatment groups using CMH test with randomization stratification factor(s) adjusted.

10.4.12 Hemoglobin Increase from Baseline ≥ 1.0 g/dL Response Based on Rolling Method

The hemoglobin response will be measured using the consecutive "rolling" 12-week and 24-week time intervals within the entire study period up to the efficacy cutoff, i.e., Days 2 to 85, Day 3 to 86 for 12-week (or Day 2 to 169, Day 3 to 170 for 24-week) and so on. Note that, Day 1 hemoglobin belongs to baseline. The hemoglobin response is defined as subjects with ≥ 1.0 g/dL increase in mean from baseline during any consecutive rolling 12-week (or 24-week) time interval in the absence of transfusion. The Hb values within 21 days following a transfusion may be influenced by the transfusion and will be excluded from this analysis.

The treatment comparison with "rolling" method (luspatercept versus placebo) will be conducted by the CMH test stratified by the randomization stratification factors. The OR (luspatercept versus placebo) with corresponding 2-sided (at 0.05 alpha level) 95% CI and p-value will be provided. The number and percentage of responders will be summarized by each treatment group and the difference in proportions (luspatercept – placebo) and corresponding 95% CI will also be calculated by exact unconditional test with "rolling" method as well.

10.4.13 Duration of Mean Hemoglobin Increase from Baseline ≥ 1.0 g/dL

To measure the duration of hemoglobin response (for the ≥ 1.0 g/dL criterion), both the 12-week and 24-week rolling methods mentioned from Section 10.4.12 will be applied.

The duration of the longest continuous 12-week (or 24-week) based hemoglobin response will start from the first day of the first rolling 12-week (or 24-week) interval that achieves mean Hb increase from baseline ≥ 1.0 g/dL with the exclusion of hemoglobin assessments within 21 days after transfusion. It will end with the last day of the last consecutive rolling 12-week (or 24-week) interval that maintains mean Hb increase from baseline ≥ 1.0 g/dL.

The Kaplan-Meier plots of duration of mean hemoglobin increase from baseline ≥ 1.0 g/dL over rolling 12-week (or 24-week) time interval, defined as time from the start of the longest response to end of the response, will be provided. The median duration, 25th and 75th quartiles with the associated 2-sided (at 0.05 alpha level) 95% CIs will be presented for each treatment group. Descriptive statistics will also be generated for the duration for each treatment group.

The duration of the individual continuous response is defined as the number of days between the First and Last Days of Response, i.e. Last Day of Response – First Day of Response + 1, where the First Day of Response is the first day of the first 12-week interval when the subject meets response requirement, and the Last Day of Response is the last day of the last 12-week (or 24-week) interval when the subject meets response requirement. The subject must meet the response requirement on all the days within the entire duration.

For subjects who have one response and continue to respond at the efficacy cutoff date, the end day of the response will be censored at the date of efficacy cutoff and the duration of response will be calculated as date of efficacy cutoff – first day of response + 1, where date of efficacy cutoff is defined in Section 10. For subjects who have multiple responses and their last responses continue on the efficacy cutoff date, the longest response will be the last one if the duration from the response start to censoring is longer than all the previously occurred response durations. If the continuing response duration is not the longest compared with the previously occurring responses, the response with the longest duration will be selected. Summary statistics will be provided for the total duration of hemoglobin response within the entire study period and the ratio of the total response duration versus entire study duration. The entire study duration is defined as period from Day 2 to date of efficacy cutoff.

10.4.14 6MWT: Mean Change at Week 24 and 48

For 6MWT assessments, mean changes from baseline at Weeks 24 and 48 will be summarized descriptively and analyzed using ANCOVA method with treatment group, randomization stratification factor(s) and baseline 6MWT value as the covariates. In addition, subgroups of subjects with baseline 6MWT ≤ 450 meters and > 450 meters will be performed. Other subgroup analyses will be added as described in Section 10.5.

Mean changes from baseline at Week 48 will be analyzed similarly.

10.4.15 Proportion of Mean Hemoglobin Increase from Baseline ≥1.5 g/dL at Week 13 to 24

Similar to the primary efficacy analysis, the proportion of subjects who have an increase from baseline ≥1.5 g/dL in mean of hemoglobin values over a continuous 12-week interval from Week 13 to Week 24 in the absence of transfusions will be performed on the ITT population. The Hb values within 21 days following a transfusion may be influenced by the transfusion and will be excluded from this analysis. Cochran-Mantel-Haenszel (CMH) test will be performed with treatment group and randomization stratification factor(s) in the model to compare the treatment and placebo groups at 2-sided 0.05 level.

10.4.16 Proportion of subjects with a decrease from baseline ≥ (RD) in mean NTDT-PRO T/W score, over Weeks 13 to 24 and Weeks 37 to 48

The responder definition (RD) threshold is the individual subject score change over a predetermined time period that will be interpreted as a treatment benefit. The RD threshold of 1 for the NTDT-PRO is jointly determined by longitudinal anchor based and distribution-based methods prior to database lock using data collected over a 12-week window between Weeks 13 to 24 ([redacted]).

Proportion of subjects with a decrease from baseline ≥ RD (=1) in mean NTDT-PRO T/W score, over Weeks 13 to 24 and Weeks 37 to 48 in the luspatercept group will be compared with the placebo group. The CMH test will be used with baseline NTDT-PRO T/W category ≥ 3 versus <3 as the stratum to compare the response rates between the 2 groups. The corresponding 95% confidence interval for odds ratio will also be provided.

The cumulative distribution function (CDF) of observed changes in NTDT-PRO T/W over a continuous period of Weeks 13 to 24 and Weeks 37 to 48 from baseline will be generated by treatment group. The CDF curve presents the entire distribution of responses and shows a continuous plot of the entire range of changes from baseline at a given post-baseline visit on the x-axis and the cumulative percentage of subjects experiencing a change up to a particular level on the y-axis.

10.4.17 Time from First Dosing Date to the First Mean Hemoglobin Response

To measure the time to the first hemoglobin response (for the ≥ 1.0 g/dL criterion) in the absence of transfusions, both 12-week and 24-week rolling methods mentioned from section 10.4.12 will be applied.

The descriptive statistics for the time from first dosing date to the first 12-week (or 24-week) mean hemoglobin increase response (for the ≥ 1.0 g/dL criterion) will be provided by treatment group, where time from first dosing date to the first Hb increase response is defined as First Day of 12-week (or 24-week) Response – Date of First Study Drug +1. The difference in time from the first dosing date to the first Hb increase response (luspatercept – placebo), corresponding 95% CI and p-value will be calculated by t-test. Only subjects who have a response will be included.

10.4.18 Longitudinal Analysis of Hemoglobin Mean Change from Baseline

A mixed effect repeated measure model will be used with intercept and time as the random effect to evaluate the hemoglobin change from baseline at Weeks 12, 24, 36 and 48. The model includes the weekly change from baseline up to Week 48, and treatment group assignment, time (weeks since Dose 1 Day 1), baseline hemoglobin, randomization stratification factor(s), time-by-baseline score, and time-by-treatment group assignment terms in the model. The unstructured covariance matrix will be used to obtain the random effects variance components.

If the MMRM model using an unstructured covariance matrix fails to converge with the default Newton-Raphson algorithm, the Fisher scoring algorithm or another appropriate method will be used to provide initial values of the covariance parameters. In the rare event that none of the above methods yields convergence, a structured covariance, such as heterogeneous Toeplitz and Toeplitz structures, will be used to model the correlation among repeated measurements. In this case, the empirical option will be used because the sandwich variance-covariance estimator is asymptotically consistent.

10.5 Subgroup Analysis

The primary and key secondary efficacy endpoints along with the FACIT-F FS score, LIC/ICT response and 6-minute walk test will also be summarized for the following subgroups:

1. Geographic region:
    - North America and Europe
    - Middle East
    - Asia-Pacific
2. Age:
    - $\leq$ 32 years
    - $>$ 32 years
3. Splenectomy:
    - Y
    - N
4. Sex:
    - Male
    - Female
5. Beta-Thalassemia diagnosis:
    - Beta-Thalassemia

- Hemoglobin E/Beta-Thalassemia
- Beta-Thalassemia combined with Alpha-Thalassemia;

6. Baseline liver iron content
   - < 3 mg/g dry weight
   - >=3 - <= 5 mg/g dry weight
   - > 5 - <= 7 mg/g dry weight
   - >7 - <= 15 mg/g dry weight
   - > 15 mg/g dry weight
7. Mean baseline Hb (mean of at least 2 Hb values by Central Lab during the 28 days screening period) (g/dL)
   - >=8.5 g/dL
   - <8.5 g/dL
8. Baseline NTDT-PRO T/W Score
   a. ≥ 3 points
   b. < 3 points
9. Baseline 6MWT
   a. ≤ 450 meters
   b. > 450 meters
10. Baseline FACIT-F FS Score
    a. ≥ 43
    b. < 43

A forest plot showing the odds ratios, 95% CI and p-value for the overall result and the results in each subgroup will be constructed for primary and key secondary endpoints.

10.6 Handling of Missing Data or Dropouts

10.6.1 Missing Data for non-QoL Assessment

In order to assess the robustness of the findings on the primary and key secondary endpoints of hemoglobin response and change from baseline, the missing values will be imputed as follows:

- For binary responder/non-responder values, the non-responder imputation (NRI) method is used for primary analysis. The NRI method classifies all subjects with missing data as non-

responders. The supportive/sensitivity analysis will include the complete case analysis. The complete case analysis excludes all subjects with missing data.

Tipping point analysis may be conducted for the primary efficacy endpoint if necessary. Different number of responders between luspatercept and placebo will be assessed until the study conclusion is changed. Each imputed value is initially imputed as a non-responder. The imputed values in the luspatercept group will remain as non-responders while the imputed values in the placebo group are replaced as a responder one at a time, therefore changing the number of responders between groups. Once all imputed values in the placebo group have been replaced with responder values, one of the imputed values in the luspatercept group is replaced with a responder while all the imputed values in the placebo group are reset to all non-responders. This process of changing each imputed non-responder to a responder one at a time continues until all the imputed values in the luspatercept and placebo groups are changed to responders.

In this way, every pair of imputations between the placebo and luspatercept groups will be assessed similarly, creating a matrix of all possible missing patterns. At each iteration, the statistical analysis is assessed to result in a significant or non-significant p-value, and the result of the analysis is recorded. A table will be provided to identify at what point to which the difference in responder status causes a directional shift in the study result.

- For continuous values, the complete case analysis is the primary analysis. The supportive/sensitivity analysis will use the postbaseline mean imputation, the baseline imputation or multiple imputation (MI) method.

In the case of any missing data for hemoglobin and LIC from MRI, imputation will be applied for each section. The imputation logic for missing LIC value is stated in Section 10.4.10: the value of LIC will be either the value collected from eCRF or the value derived from T2*, R2* or R2 parameters depending on which techniques and software were used for MRI LIC data acquisition.

10.6.2 Missing Data for QoL Assessment

For continuous and categorical QoL assessments, the complete case analysis is used for the primary analysis. For the key secondary endpoint, NTDT-PRO T/W domain score at Week 13-24, the supportive/sensitivity analysis will use the postbaseline mean imputation, the baseline imputation or multiple imputation (MI) methods to assess the robustness of the primary result.

To assess the extent of missing data at each assessment visit by treatment group, compliance rates for the NTDT-PRO, FACIT-F, and SF-36 assessments at each scheduled visit will be provided. The compliance rate for each assessment at each visit is defined as follows:

- NTDT-PRO: The number of subjects with at least half of 6 weekly item scores for that week being non-missing (i.e., ≥ 3 items) divided by the number of ITT subjects who are still expected to provide PRO assessment at that time point.

- FACIT-F: The number of subjects with at least 80% of the 40 items (i.e., ≥ 32 items) for that visit being completed divided by the number of ITT subjects who are still expected to provide FACIT-F assessment at that time point.

- SF-36: The number of subjects with at least 50% of the 36 items (i.e., ≥ 18 items) for that visit being completed divided by the number of ITT subjects who are still expected to provide SF-36 assessment at that time point.

11 SAFETY ANALYSIS

The purpose of this section is to define the safety parameters for the study. All summaries of safety data will be conducted using the safety population. The safety analysis includes adverse events (AEs), clinical laboratory tests, vital signs, electrocardiogram (ECG), echocardiography, erythropoietin, left ventricular ejection fraction (LVEF), and antidrug antibody (ADA) testing. In addition, pregnancy test and menstrual status assessments will be provided for female subjects.

If not otherwise specified (for example, adverse events), safety summaries will use all collected records.

11.1 Adverse Events

Adverse events will be analyzed in terms of treatment-emergent adverse events (TEAEs) which are defined as any AEs that begin or worsen on or after the start of study drug through 63 days after the last dose of IP (i.e., AE start date on or after the first dose date and within last dose date + 63). In addition, an AE that occurs beyond this timeframe and that is assessed by the investigator as possibly related (suspected) to study drug will be considered as treatment-emergent.

All AEs will be coded using the Medical Dictionary for Regulatory Affairs® (MedDRA) dictionary. The incidence of TEAEs will be summarized by MedDRA system organ class (SOC) and preferred term (PT). The AE tables will be sorted by SOC and PT (within SOC) in descending frequency within the luspatercept group. If a subject experiences multiple AEs under the same PT (or SOC), then the subject will be counted only once for that PT (or SOC).

The intensity of AEs will be graded 1 to 5 according to the National Cancer Institute Common Terminology Criteria for Adverse Events (CTCAE) Version 4.0. If a subject experience the same AE more than once with different toxicity grades, then the event with the highest grade will be tabulated in "by grade" tables. In addition, AEs with a missing intensity will be presented in the summary table as an intensity category of "Missing" only if the same event category has no other valid grades.

Tables summarizing the incidence of TEAEs will be generated for each of the following by treatment group (if not otherwise specified, the summary is by SOC and PT):

- TEAEs;
- TEAEs by SOC only;
- Treatment-related TEAEs;
- Serious TEAE;
- Treatment-related serious TEAEs;
- TEAEs by CTCAE maximum severity
- Treatment-related TEAEs by CTCAE maximum severity
- TEAE with CTCAE Grade ≥ 3;

- Treatment-related TEAE with CTCAE Grade ≥ 3;
- TEAEs leading to study drug discontinuation;
- Treatment-related TEAEs leading to study drug discontinuation;
- TEAEs leading to study drug dose reduction;
- Treatment-related TEAEs leading to study drug dose reduction;
- TEAEs leading to death;
- Treatment-related TEAEs leading to death;
- Most frequent TEAEs by PT (≥ 5% in PT frequency of all subjects or subjects from any treatment group);
- Most frequent TEAEs by SOC (≥ 5% in SOC frequency of all subjects);
- Most frequent TEAEs by SOC and high level term (HLT) (≥ 5% in HLT frequency of all subjects);
- TEAEs by Age group (≤ 32 years, > 32 years);
- TEAEs by Age group (≤ 32 years, 32-≤ 50 years, > 50 years);
- TEAEs by Gender (male and female);
- TEAEs by Splenectomy status (Yes/No);
- TEAEs by ADA status (positive preexisting, positive treatment-emergent, negative);
- TEAE by preferred term for QTC Prolongation and Atrial Fibrillation events;
- All death by cause of death;

Listings for AEs, SAE, and AEs leading to drug discontinuation will be presented separately. Treatment-emergent AEs will be flagged in the listings. A death listing will be provided for all death events.

11.2 Adverse Events of Special Interest

The following adverse events are of special interest:

- Malignancy
- Premalignant Conditions.

An AESI summary by PT terms will be provided for TEAEs by treatment group. A listing for AESI will be provided as support.

To evaluate the timing of AESI onset, descriptive statistics will be provided for time to first AESI occurrence after treatment by treatment group A listing for AESI will be provided as support.

11.3 Other Adverse Events That Require Safety Analysis

Other adverse events that require safety analysis include AEs that fall under the "Embolic and Thrombotic Events and Thrombophlebitis" SMQ category and AEs with PT "Bone pain" and 'Hypertension'.

Similar to AESI, summary will be provided for AEs that fall under the "Embolic and Thrombotic Events and Thrombophlebitis" SMQ category by preferred term. For "Bone pain" events, number of subjects with bone pain events by worst CTCAE grade will be provided from grade 1 to grade 5. The number and percentage of subjects with bone pain occurred during the first 24 weeks and after 24 weeks will be provided. Summary statistics will also be provided for time to first bone pain, total duration of bone pain (in days), total duration of bone pain that occurred during the first 24 weeks (in days), and total duration of bone pain that occurred after 24 weeks (in days). Total duration of bone pain is defined as sum of all bone pain duration within a subject, excluding overlapped period. Same summaries are also conducted on 'Hypertension'.

The "Embolic and Thrombotic Events and Thrombophlebitis" SMQ category analysis will include subgroup analysis by splenectomy status, platelet above ULN, concomitant medications, and comorbidities as needed.

11.4 Clinical Laboratory Evaluations

Clinical laboratory data is collected by central lab and local lab. Lab data will be collected over time during the study. All summaries will be based on the SI units and missing values will not be imputed. Clinical laboratory values will be graded (grade 0-4) according to NCI-CTCAE version 4.0 for applicable tests. Normal ranges will be used to determine the "High", "Low", and "Normal" categories for all numeric laboratory tests. Only central lab results and local lab "Reticulocyte (Blood)" parameter will be used for table summaries.

11.4.1 Hematology/Chemistry/Immunology

The laboratory results and change from baseline will be summarized by visit by treatment group for central lab hematology and chemistry panels separately.

A shift table representing the shift from the baseline grade to maximum NCI-CTC Grades (high or low) will be provided for selected hematology and chemistry parameters having toxicity grade by visit by treatment group. The shift summary for high category will be done for hepatic function parameter (ALT, AST, ALP, total bilirubin) and renal function parameter (serum creatinine and eGFR by MDRD). The shift summary for low category will be done for hematology parameter (platelets, leukocytes and absolute neutrophil counts).

To estimate the incidence of subjects who have passed the predefined threshold for selected parameters, a summary table representing the number and percentage of subjects with lab assessments satisfying the threshold criteria will be provided by treatment group. A subject with post-baseline result (including "unscheduled" visits) meeting the criteria will be counted. The threshold criteria include below:

LIVER FUNCTION

Post-baseline Alanine Aminotransferase (ALT)	≥ 3x upper limit of normal (ULN)
Post-baseline Aspartate Aminotransferase (AST)	≥ 3x ULN
Post-baseline Total Bilirubin (BILTOTAL)	≥ 2x ULN
Post-baseline ALT/AST and BILTOTAL	(ALT≥ 3x ULN or AST≥ 3x ULN) and BILTOTAL≥ 2x ULN
RENAL FUNCTION	
Post-baseline eGFR by MDRD (CREATCLR)	< 0.5x baseline
Post-baseline Serum Creatinine (CREAT)	> 2x baseline
Albuminuria Category (ACR: mg/g)	<30
	≥30- ≤300
	>300-≤1000
	>1000-≤3500
	>3500
HEMATOLOGY	
Post-baseline Leukocytes (WBC)	≥ 2x baseline and > ULN
	≥ 3x baseline and > ULN
	≥ 2x baseline and > ULN and lasts for at least 42 days
	≥ 3x baseline and > ULN and lasts for at least 42 days
Maximum Post-baseline Platelets (PLAT)	≥ 1.5x baseline and > ULN
	≥600 - <1000x10^9/L
	≥ 1000 x 10^9/L

Specifically, the threshold summary for platelets will be based on the maximum post-baseline value. The summary will be provided by baseline category based on normal range (ie, within normal limit at baseline, > upper limit at baseline), and by splenectomy status (Yes, No) in separate tables. Furthermore, the number and percentage of subjects with maximum post-baseline WBC exceeding 3x baseline value and >ULN, maximum post-baseline WBC exceeding 3x baseline value and > ULN and lasts for at least 42 days, and subjects with maximum post-baseline platelets ≥600 - <1000x10^9/L and maximum post-baseline platelets exceeding 1000 x 10^9/L will be provided separately by splenectomy status (Yes, No).

For some key lab parameters (ALT, AST, WBC), plots will be presented to show the pattern of the lab test values over time by treatment group. Mean and SE will be presented in the plot.

Listings of clinical laboratory data will be provided for central lab and local lab respectively for each panel in safety analysis (excluding serum erythropoietin and serum ferritin). Abnormal observations will be noted. Specifically, subjects with any WBC differential count exceeding 2x baseline value will be listed in a separate listing. All WBC differential count records of qualified subjects will be presented.

11.4.2 Serum Erythropoietin

Serum erythropoietin are collected from central laboratory. The summary of serum erythropoietin test results and change from baseline will be provided by visit by treatment group. A plot will be presented to show the pattern of the serum erythropoietin test results over time by treatment group. Mean and SE will be presented in the plot. A listing will be provided.

11.4.3 Local lab "Reticulocyte (Blood)" parameter

The "Reticulocyte (Blood)" parameter is only collected at local lab. The summary of absolute reticulocyte count and change from baseline will be provided by visit and treatment group in the same way as other lab parameters.

A line plot will be presented to show the pattern of the reticulocyte test results over time by treatment group. Mean and SE will be presented in the plot.

11.5 Vital Sign Measurements

Vital sign is collected over time during the study. Vital sign parameters include weight, temperature, pulse rate, seated blood pressure (diastolic blood pressure [DBP] and systolic blood pressure [SBP]). The DBP and SBP are collected twice at each visit with 10 minutes apart. The average of the two assessments will be used. Summary statistics of observed values and change from baseline values will be presented for each parameter by visit by treatment group.

To further estimate the incidence of subjects whose maximum post-baseline blood pressure have passed selected criteria, summary tables representing the number and percentage of subjects with post-baseline (including 'unscheduled' visits) SBP/DBP assessments satisfying each criterion will be provided by treatment group. The selected criteria include below:

Maximum post-baseline SBP	No increase
	Increased < 20 mmHg
	Increased ≥ 20 mmHg
	Increased ≥ 20 mmHg and SBP ≥ 140 mmHg
	Increased ≥ 20 mmHg and SBP ≥ 150 mmHg
	Subjects only with baseline values
Maximum post-baseline DBP	No increase

	Increased < 20 mmHg
	Increased ≥ 20 mmHg
	Increased ≥ 20 mmHg and DBP ≥ 100 mmHg
	Subjects only with baseline values

A plot will be presented to show the pattern of the SBP and DBP test results over time by treatment group. Mean and SE will be presented in the plot. Additionally, a spaghetti plot for SBP and DBP values over time for individual subjects with maximum post-baseline SBP increased ≥ 20 mmHg and SBP ≥ 150 mmHg or maximum post-baseline DBP increased ≥ 20 mmHg and DBP ≥ 100 mmHg will be provided. Corresponding listing will be provided for vital sign data.

11.6 Electrocardiograms

The 12-lead ECG is collected over time at selected visits. ECG parameters include heart rate, PR interval, QRS duration, RR interval and QT interval. The RR interval value will be derived per formula: RR interval (msec)=60000 (msec)/heart rate (bpm). The corrected value for QT interval will be derived based on Fridericia's formula as below:

Fridericia's formula: $QTcF = QT/(RR)^{1/3}$

where RR is the calculated RR interval as above

The calculated RR interval value, recorded values of ECG parameters and change from baseline values will be summarized by visit by treatment group.

To further estimate the incidence of subjects whose baseline or post-baseline QTcF values have passed the selected ICH E14 Criteria, summary tables representing the number and percentage of subjects with ECG assessments satisfying the CPMP (Committee for Proprietary Medicinal Products) criteria will be provided separately for QTcF by treatment group and visit (for baseline and post-baseline respectively). A subject with baseline or any post-baseline (including 'unscheduled' visits) result meeting individual criteria will be counted. The selected CPMP criteria includes below:

QTcF Interval	> 450 msec
	> 480 msec
	> 500 msec
QTcF Interval Increase from Baseline	≥ 30 msec
Post-baseline QTcF Interval and Increase from	Post-baseline Interval > 480 msec and

Baseline	Increase from Baseline >= 60 msec

Corresponding listing will be provided for ECG data.

11.7 Left Ventricular Ejection Fraction (LVEF)

The Left ventricular ejection fraction (LVEF) are collected over time at selected visits. It will be measured by either echocardiography (ECHO) or MRI. Recorded values of LVEF and change from baseline values will be summarized by treatment group and by visit.

Corresponding listing will be provided for LVEF.

11.8 Antidrug Antibody Testing

The anti-drug antibody (ADA) binding to luspatercept (anti-ACE-536) will be tested over time in all subjects. The ADA level in samples confirmed positive to anti-ACE-536 will be semi-quantitated and reported as titer. Specificity tests and neutralizing ADA tests will be conducted only for samples confirmed positive to anti-ACE-536.

To evaluate the treatment-emergent ADA level, the number and percentage of the subjects for each ADA test will be presented by treatment group and ADA status. The ADA status of a subject during treatment is determined based on the longitudinal anti-ACE-536 results as following:

- Negative: All samples (baseline and post-baseline) are negative.
- Positive to treatment-emergent ADA:
  - At least one post-baseline sample is positive if the baseline sample is negative, or
  - At least one post-baseline sample is positive with a titer ≥ 4-fold of the baseline titer if the baseline sample is positive
- Positive to preexisting ADA:
  - Baseline sample is positive and all post-baseline samples are negative, or
  - Both baseline and post-baseline samples are positive, but all positive post-baseline sample have a titer < 4-fold of the baseline titer.

A separate table will summarize the number and percentage of the subjects and median (min, max) titer for anti-ACE-536 by treatment group, visit, and ADA status. In this table, only subjects tested positive to anti-ACE-536 will be included, and only positive subjects in the luspatercept arm will be summarized by ADA status. Corresponding listing will be provided to support the table.

Additionally, a bar plot will be provided for subject's ADA status ("Preexisting", "Treatment-Emergent" and "Negative"). Percent of subjects within each category will be displayed by treatment group.

11.9 Pregnancy Test and Menstrual Status for Female Subjects

The number and percentage of the subjects for each pregnancy test result category (i.e., positive, negative) will be presented by treatment group. A subject is counted as 'positive' if there is any positive result captured after first dose date, a subject is counted as 'negative' if there is no positive result captured after first dose date.

The pregnancy test along with menstrual status will be provided in a listing.

14 HEALTH RELATED QUALITY OF LIFE ANALYSIS

The HRQoL analyses are addressed in corresponding subsection in Section 10.4. This SAP only provides description of main HRQoL analysis specify in the protocol. A detailed statistical analysis of HRQoL data will be provided in a separate HRQoL SAP, which will be finalized prior to database lock. The HRQoL SAP will be appended to separate specific HRQoL report.

15 GENERAL INFORMATION

15.1 Interim Analysis

There is no interim analysis planned for this study.

15.2 DMC

The independent DMC will be composed of experts in the β-thalassemia not involved in ACE-536-B-THAL-002 protocol, an Independent Cardiologist, an Independent Statistician, and may include additional ad hoc members. Representatives of Sponsor will be attending the blinded part of the DMC meetings. The Sponsor will not have access to the unblinded data.

During the course of the study, the DMC will review unblinded safety data regularly, or ad-hoc, as well as safety and efficacy data in accordance with the guidelines for the preplanned analyses.

An independent third party will prepare the reports of aggregate data summaries and individual subject data listings, as appropriate, to the DMC members for each scheduled or ad-hoc meeting.

The DMC responsibilities, authorities, and procedures will be detailed in the DMC charter, which will be endorsed by the DMC prior to the first data review meeting.

Operational details for the DMC and the algorithm will be detailed in the DMC charter.

16 IMPACT OF COVID-19 ON EFFICACY AND SAFETY ANALYSIS

The outbreak of a respiratory disease caused by a novel coronavirus has quickly become a global pandemic. The virus has been named "SARS-CoV-2" and the disease it causes has been named "Coronavirus Disease 2019" (COVID-19). When the pandemic started, the study has finished enrollment of all patients and is on the way to collect the clinical data for the last 6 months.

The COVID-19 pandemic may impact the conduct and statistical analysis of clinical trials in 3 different aspects ([redacted]

1. Indirect (operational) impact – quarantine/travel restrictions, site closure, interruption of supply chain to investigational product, overwhelmed healthcare systems, enrollment slow/pause.
2. Direct impact on trial participants - COVID-19 infection, treatment for COVID-19.
3. Impact that may affect endpoint interpretation - delayed/missed visits/assessments, treatment delayed/interrupted/discontinued, study withdrawal, alternative ways of treatment administration, alternative ways of data collection.

Any subjects who had any pandemic-related scenarios mentioned above will be defined as the COVID-19-impacted population in the pandemic period. Any subjects who had contracted COVID-19 will be grouped as the COVID-19-infected population. Therefore, the COVID19-infected population is a subgroup of the COVID19-impacted population. The identification of the COVID-19-impacted and COVID-19-infected patients will be conducted on a weekly base and will be finalized prior to the database lock.

COVID-19 information for the COVID-19-impacted population will be captured in tables of subject disposition, concomitant medication, healthcare resource utilization, summary of dose delay, and protocol deviation. All these 5 tables will have the COVID-19-related fields. The subject disposition and dose delay tables provide the cause of discontinuation due to COVID-19. The protocol deviation table have sections specifying whether the deviations were caused by COVID-19. The table of concomitant medication captures COVID-19 medication if a subject is infected. The healthcare resource utilization table includes the reason for hospitalization caused by COVID-19.

16.1 Sensitivity Analysis of COVID-19 Impact on Efficacy Endpoints

The sensitivity analyses may be conducted for the primary and key secondary efficacy endpoints of the non-COVID-19-impacted patients by removing the COVID-19-impacted population to evaluate the effects of dose delays, missed central lab measurements, missed assessments, COVID-19 infection, or early discontinuation due to COVID-19.

16.2 Analysis and Reporting of COVID-19 Impact on Safety Endpoints

The summary analyses for safety endpoints will be conducted on COVID-19-infected populations for any adverse events during the entire pandemic period.

Tables summarizing the incidence of TEAEs will be generated for each of the following by treatment group (if not otherwise specified, the summary is by SOC and PT) for COVID-19-infected populations:

- TEAEs;
- Serious TEAE (SAE);
- TEAE with CTCAE Grade ≥ 3;
- TEAEs leading to study drug discontinuation.

18 APPENDICES

18.1 Handling of Dates

Dates will be stored as numeric variables in the SAS analysis files and reported in DDMMMYYYY format (ie, the Date9. datetime format in SAS). Dates in the clinical database are classified into the categories of procedure dates, log dates, milestone dates, and special dates.

- **Procedure Dates** are the dates on which given protocol-specified procedure are performed. They include the dates of laboratory testing, physical examinations, tumor scans, etc. They should be present whenever data for a protocol-specified procedure are present and should only be missing when a procedure is marked as NOT DONE in the database. Procedure dates will not be imputed.
- **Log Dates** are dates recorded in eCRF data logs. Specifically, they are the start and end dates for adverse events and concomitant medications/procedures. They should not be missing unless an event or medication is marked as *ongoing* in the database. Otherwise, incomplete log dates will be imputed according to the rules in Appendix 18.2 (eg, for duration or cycle assignment, etc.). However, in listings, log dates will be shown as recorded without imputation.
- **Milestone Dates** are dates of protocol milestones such as randomization, study drug start date, study drug termination date, study closure date, etc. They should not be missing if the milestone occurs for a subject. They will not be imputed.
- **Special Dates** cannot be classified in any of the above categories and they include the date of birth. They may be subject to variable-specific censoring and imputation rules.

Dates recorded in comment fields will not be imputed or reported in any specific format.

18.1.1 Calculation Using Dates

Calculations using dates (eg, subject's age or relative day after the first dose of study drug) will adhere to the following conventions:

- Study days after the start day of study drug will be calculated as the difference between the date of interest and the first date of dosing of study drug plus 1 day. The generalized calculation algorithm for relative day is the following:
  - If TARGET DATE ≥ DSTART then STUDY DAY = (TARGET DATE – DSTART) + 1;
  - Else use STUDY DAY = TARGET DATE – DSTART.

  Note that Study Day 1 is the first day of treatment of study drug. Negative study days are reflective of observations obtained during the baseline/screening period. Note: Partial dates for the first study drug are not imputed in general. All effort should be made to avoid incomplete study drug start dates.

- Age (expressed in years) is calculated: the number of months between birth date and informed consent date divided by 12 (if both dates are not missing), the integer part will be kept. If the month of birth date is the same as informed consent date and the day of birth date is greater than informed consent date, then the age calculated by above will minus 1. If any date is missing, AGE will be set to the age collected from CRF.
  - Partial birth date: impute missing day as 15th of the month; impute missing month as July; set missing age for missing year
- Intervals that are presented in weeks will be transformed from days to weeks by using (without truncation) the following conversion formula:

  WEEKS = DAYS /7
- Intervals that are presented in months will be transformed from days to months by using (without truncation) the following conversion formula:

  MONTHS = DAYS /30.4167

18.2 Date Imputation Guideline

18.2.1 Impute Missing Dates of Adverse Events/ Prior or Concomitant Medications, Procedures/Surgeries

Incomplete Start Date:

Missing day and month

- If the year is the **same** as the year of the first dosing date, then the day and month of the first dosing date will be assigned to the missing fields.
- If the year is **prior to** the year of first dosing date, then December 31 will be assigned to the missing fields.
- If the year is **after** the year of first dosing, then January 1 will be assigned to the missing fields.

Missing day only

- If the month and year are the **same** as the year and month of first dosing date, then the first dosing date will be assigned to the missing day.
- If either the year of the partial date is **before** the year of the first dosing date or the years of the partial date and the first dosing date are the same but the month of partial date is **before** the month of the first dosing date, then the last day of the month will be assigned to the missing day.
- If either the year of the partial date is **after** the year of the first dosing date or the years of the partial date and the first dose date are the same but the month of partial date is **after** the month of the first dosing date, then the first day of the month will be assigned to the missing day.

- If the stop date is not missing, and the imputed start date is after the stop date, the start date will be imputed by the stop date.

Missing day, month, and year

- No imputation is needed, the corresponding AE will be included as TEAE.

Incomplete Stop Date: If the imputed stop date is before the start date, then the imputed stop date will be equal to the start date.

Missing day and month

- If the year of the incomplete stop date is the **same** as the year of the last dosing date, then the day and month of the last dosing date will be assigned to the missing fields.
- If the year of the incomplete stop date is **prior to** the year of the last dosing date or prior to the year of the first dosing date, then December 31 will be assigned to the missing fields.
- If the year of the incomplete stop date is **prior to** the year of the last dosing date but is the same as the year of the first dosing date, then the first dosing date will be assigned to the missing date.
- If the year of the incomplete stop date is **after** the year of the last dosing date, then January 1 will be assigned to the missing fields.

Missing day only

- If the month and year of the incomplete stop date are the **same** as the month and year of the last dosing date, then the day of the last dosing date will be assigned to the missing day.
- If either the year of the partial date is **not equal to** the year of the last dosing date or the years of the partial date and the last dosing date are the same but the month of partial date is **not equal to** the month of the last dosing date, then the last day of the month will be assigned to the missing day.

18.3 QoL Algorithm

18.3.1 NTDT-PRO (Non-Transfusion-Dependent Thalassemia Patient-Reported Outcome)

The NTDT-PRO is a disease-specific QoL tool used to assess the severity of symptoms associated with NTDT β-thalassemia. It is administered as a daily eDiary with recall of thalassemia-related symptoms during the past 24 hours.

Baseline and post-baseline score calculations are performed based on weekly item scores. Baseline scores are calculated over 7 days prior to Dose 1 Day 1. The NTDT-PRO scores are recorded daily from Dose 1 Day 1 through Week 24. After Week 24, the NTDT-PRO scores are recorded by visit. The Week 1 weekly scores are calculated from Day 1 to Day 7. The Week 2 weekly scores are calculated from Day 8 to Day 14, similarly through the Week 24 weekly scores. After Week 24, the weekly scores will be derived over 7 days based on the closest visit day.

Step 1: Scoring the Items

All items have score ranges from 0–10, with higher scores indicating more severe symptoms. If entries are missing for ≥ 4 days in any week, the average weekly item score is declared as missing. Otherwise the mean of the non-missing item values is taken.

If subjects do not have any valid NTDT-PRO scores at baseline, Week 13-24 or Week 37-48 due to the minimum 4-day non-missing weekly score requirement, the 4-day non-missing rule will be relaxed to at least 2 days to avoid missing data. The justification of this relaxed rule is supported by the simulation results showing that no significant difference is found between 4-day and 2-day non-missing weekly scores (*Psychometric assessment of the NTDT-PRO using data from the phase 2 luspatercept study, 2020*).

Step 2: Calculating Scores for the T/W Domain

The average of non-missing weekly item scores from the first 4 items (i.e., TiredNA, TiredPA, WeakNA, WeakPA) will be calculated if at least 1 non-missing weekly item score for tiredness [i.e. TiredNA, TiredPA] and at least 1 non-missing weekly item score for weakness [i.e. WeakNA, WeakPA]; otherwise, the weekly T/W score will be set to missing. Thus, the range of weekly T/W score is 0–10.

Step 3: Calculating Scores for the SoB Domain

The average of non-missing weekly item scores from the last 2 items (i.e., SobNA and SobPA) will be calculated if ≥50% of the items (≥1 items) are not missing; otherwise, the weekly SoB score will be set to missing. Thus, the range of weekly SoB score is 0–10.

NTDT-PRO© (version 2) Instrument

For each of the following questions, please choose the number that best describes the symptoms that you may have experienced during the **past 24 hours**.

1. How would you rate your tiredness (lack of energy) when you were not doing physical activity during the past 24 hours?

☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐

0 1 2 3 4 5 6 7 8 9 10

No tiredness — Extreme tiredness

2. How would you rate your tiredness (lack of energy) when you were doing physical activity during the past 24 hours?

☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐

0 1 2 3 4 5 6 7 8 9 10

No tiredness — Extreme tiredness

3. How would you rate your weakness (lack of strength) when you were not doing physical activity during the past 24 hours?

☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐

0 1 2 3 4 5 6 7 8 9 10

No weakness — Extreme weakness

4. How would you rate your weakness (lack of strength) when you were doing physical activity during the past 24 hours?

☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐

0 1 2 3 4 5 6 7 8 9 10

No weakness — Extreme weakness

5. How would you rate your shortness of breath when you were <u>not</u> doing physical activity during the past 24 hours?

☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
0	1	2	3	4	5	6	7	8	9	10
No shortness of breath										Extreme shortness of breath

6. How would you rate your shortness of breath when you were doing physical activity during the past 24 hours?

☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
0	1	2	3	4	5	6	7	8	9	10
No shortness of breath										Extreme shortness of breath

For validation purposes only

7. During the past 24 hours, how would you rate the overall severity of your thalassemia symptoms?

☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
0	1	2	3	4	5	6	7	8	9	10
No symptoms										Very severe symptoms

<u>For validation purposes only (administered once every 3 weeks):</u>

8. How would you rate the overall change in your thalassemia symptoms since the start of this study?

- ☐ A great deal better
- ☐ Much better
- ☐ A little better
- ☐ No change
- ☐ A little worse
- ☐ Much worse
- ☐ A great deal worse

18.3.2 FACIT-F

FACIT-F evaluation include FACT-G scale and FACIT Fatigue Subscale (FS). FACT-G scale are based on 4 domains: Physical Well-Being (PWB), Social/Family Well-Being (SWB), Emotional Well-Being (EWB), and Functional Well-Being (FWB). In ERT system, the questions of Fatigue scale were in the section of Additional Concern. The version 4 of FACIT-F was applied in the study.

Domain	Questions
PWB	1, 2, 3, 4, 5, 6, 7
SWB	8, 9, 10, 11, 12, 13, 14
EWB	15, 16, 17, 18, 19, 20
FWB	21, 22, 23, 24, 25, 26, 27
FS	28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40

The 5 domain scores and the total score are obtained by the following steps:

Step 1: Score the Individual Question

All questions have score range 0-4 as collected and higher score indicates better QoL. So some of item response, ie raw collection need be reversed by subtracting 4 to obtain the item score. The score will be reversed from response for following questions:

Domain	Questions
PWB	1, 2, 3, 4, 5, 6, 7
EWB	15, 17, 18, 19, 20
FS	28, 29, 30, 31, 32, 33, 36, 37, 38, 39, 40

For the other questions, the response is the question score.

Step 2: Calculating Domain Score

If more than 80% of the total questions in the domain are answered, the domain score will be calculated as

$$\frac{\text{Sum of item score in the domain}}{\text{Number of items answered in the domain}} \times \text{Number of items in domain}$$

Otherwise, it will be left as missing.

Step 3: Calculating Total Score

If all domain scores are derived, the total score is the sum of all domain scores

$$\text{FACIT-F Total Score} = \text{PWB score} + \text{SWB score} + \text{EWB score} + \text{FWB score} + \text{FS score}$$

Otherwise, it will be left as missing.

18.3.3 SF-36

The SF-36 is a self-administered instrument consisting of 8 multi-item scales that assess 8 health domains: (1) Physical functioning (PF), 10 items from 3a to 3j, (2) Role-Physical (RP), 4 items from 4a to 4d, (3) Bodily Pain (BP), items 7 and 8, (4) General Health (GH), items 1 and 11a to 11d, (5) Vitality (VT), items 9a, 9e, 9g, and 9i, (6) Social functioning (SF), items 6 and 10, (7) Role-Emotional (RE), items 5a, 5b, and 5c, (8) Mental Health (MH), 5 items 9b, 9c, 9d, 9f and 9h. Two summary scales, Physical Component Summary (PCS) and Mental Component Summary (MCS), will be calculated using norm-based scores from the 8 health domains. The primary interests of the SF-36 are the 8 health domain scores and the PCS and MCS scores.

The 8 health domain scores and the PCS and MCS scores are obtained by the following steps.

Step 1: Score the 8 Health Domain

All questions will be scored as per the raw data value collected with the following exceptions. Seven questions (6, 9a, 9d, 9e, 9h, 11b, and 11d) will use the reversed score, so that 5=1, 4=2, 3=3, 2=4 and 1=5. The higher individual score indicates the better health.

The SF-36 scoring system relies on an assumption of linearity among the responses. However, question 1, 7 and 8 require recalibration to satisfy this important scaling assumption.

Question 1

Q1 (verbatim responses)	Q1 (raw value)	Q1 (final response value)
Excellent	1	5.0
Very good	2	4.4
Good	3	3.4
Fair	4	2.0
Poor	5	1.0

Question 7

Q7 (verbatim responses)	Q7 (raw value)	Q7 (final response value)
None	1	6.0
Very mild	2	5.4
Mild	3	4.2
Moderate	4	3.1
Severe	5	2.2
Very severe	6	1.0

Question 8 will have its score inversed too, but also depends on the response given for question 7, in the following manner:

Q7	Q8	Q8	Q8

(raw value)	(verbatim responses)	(raw value)	(final response value)
1	Not at all	1	6
2, 3, 4, 5 or 6	Not at all	1	5
any	A little bit	2	4
any	Moderately	3	3
any	Quite a bit	4	2
any	Extremely	5	1

If question 7 is not answered then question 8 will have its score recoded to preserve linearity, in the following manner:

Q7 (raw value)	Q8 (verbatim responses)	Q8 (raw value)	Q8 (final response value)
missing	Not at all	1	6.0
missing	A little bit	2	4.75
missing	Moderately	3	3.5
missing	Quite a bit	4	2.25
missing	Extremely	5	1.0

Missing data will be handled per the SF-36 version 2 scoring guideline. If more than half of the items are answered in a multi-item domain, then the missing item(s) will be imputed as the average score of the completed items in the same domain. If no more than half of the items are answered, the missing item will be left as missing

Step 2: Calculating Domain Score and Transforming Domain Score to 0-100 Scale Score

The total raw score for each heath domain scale will be computed. The total raw score is the simple algebraic sum of the final response values for all items in a given scale. The resulting total raw score for each health domain is then transformed to a 0-100 scale score:

Health Domain	Sum of final response values	Transformed 0-100 Scale Score
PF	Sum_PF=3a+3b+3c+3d+3e+3f +3g+3h+3i+3j	T_PF = ((Sum_PF-10)/40)*100
RP	Sum_RP=4a+4b+4c+4d	T_RP = ((Sum_RP - 4)/16)*100
BP	Sum_BP=7+8	T_BP = ((Sum_BP - 2)/10)*100
GH	Sum_GH=1+11a+11b+11c+11d	T_GH = ((Sum_GH - 5)/20)*100
VT	Sum_VT=9a+9e+9g+9i	T_VT = ((Sum_VT - 4)/16)*100
SF	Sum_SF=6+10	T_SF = ((Sum_SF - 2)/8)*100
RE	Sum_RE=5a+5b+5c	T_RE = ((Sum_RE - 3)/12)*100
MH	Sum_MH=9b+9c+9d+9f+9h	T_MH = ((Sum_MH - 5)/20)*100

Step 3: Computing Norm-based Domain Scores

In order for one health domain scale be meaningfully compared with those from the other scales and that domain scores have a direct interpretation in relation to the distribution of scores in U.S. general population, the 0-100 scale score for each health domain will be converted to norm-based scores using a T-score transformation and higher norm-based score indicates better heath:

Health Domain	Transformation of 0-100 scores to z-scores	Transformation of z-scores to norm-based scores
PF	Z_PF = (T_PF - 83.29094)/23.75883	N_PF = 50 + (Z_PF*10)
RP	Z_RP = (T_RP - 82.50964)/25.52028	N_RP = 50 + (Z_RP*10)
BP	Z_BP = (T_BP - 71.32527)/23.66224	N_BP = 50 + (Z_BP*10)
GH	Z_GH = (T_CH - 70.84570)/20.97821	N_GH = 50 + (Z_GH*10)
VT	Z_VT = (T_VT - 58.31411)/20.01923	N_VT = 50 + (Z_VT*10)
SF	Z_SF = (T_SF - 84.30250)/22.91921	N_SF = 50 + (Z_SF*10)
RE	Z_RE = (T_RE - 87.39733)/21.43778	N_RE = 50 + (Z_RE*10)
MH	Z_MH=(T_MH - 74.98685)/17.75604	N_MH = 50 + (Z_MH*10)

Step 4: Computing the Aggregated Physical and Mental Component Scores

Using the standardized domain scores (z-scores), aggregated PCS and MCS score will be computed based on the weights from the U.S. general population:

Aggregate PCS score = (Z_PF * 0.42402) + (Z_RP * 0.35119) + (Z_BP * 0.31754)+ (Z_GH * 0.24954) + (Z_VT * 0.02877) + (Z_SF *-.00753) + (Z_RE *-.19206) + (Z_MH *-0.22069)

Aggregate MCS score = (Z_PF * -0.22999) + (Z_RP * -0.12329) + (Z_BP * -0.09731) + (Z_GH * -0.01571) + (Z_VT * 0.23534) + (Z_SF *0.26876) + (Z_RE *-.43407) + (Z_MH *0.48581)

Step 5: Calculating the Norm-based PCS and MCS scores:

Aggregate PCS and MCS scores are standardized using a linear T-score transformation with a mean of 50 and a standard deviation of 10 and higher component score indicates better heath:

Norm-based PCS = 50 + (Aggregate PC score * 10)

Norm-based MCS = 50 + (Aggregate MC score * 10)