Statistical Analysis Plan

A Double-Blind, Placebo-Controlled, Multicenter Study of Sirukumab as AdjunctiveTreatment to a MonoAminergic antidepressant in Adults with Major Depressive Disorder

Protocol CNTO136MDD2001; Phase 2a

CNTO136 (Sirukumab)

Final

Issue/Report Date: 03 Dec 2018
Document No.:

Compliance: The study described in this report was performed according to the principles of Good Clinical Practice (GCP).

TABLE OF CONTENTS

LIST OF IN-TEXT TABLES

ABBREVIATIONS

AE	adverse event
BMI	body mass index
CI	confidence interval
CGI-S	Clinical Global Impression - Severity
CNS	central nervous system
C-SSRS	Columbia-Suicide Severity Rating Scale
CTQ	Childhood Trauma Questionnaire
DB	double-blind
DRC	Data Review Committee
ECG	electrocardiogram
eCRF	electronic case report form
EOT	end of trial
FACIT Fatigue	functional assessment of chronic illness therapy-fatigue
FDA	Food and Drug Administration
FU	Follow-up
$HDRS_{17}$	Hamilton Depression Rating Scale
hsCR	high sensitivity C-Reactive Protein
ICH	International Conference on Harmonization
IDS-C30	Inventory of Depressive Symptomatology – Clinician Rated 30 Item Scale
ITT	Intent-to-treat
LOCF	last observation carried forward
LR	likelihood ratio
MedDRA	Medical Dictionary for Regulatory Activities
MDD	Major Depressive Disorder
MMRM	mixed-effects model using repeated measures
PHQ-9	patient health questionnaire
SAE	serious adverse event
SAP	Statistical Analysis Plan
SD	standard deviation
SHAPS	Snaith–Hamilton pleasure scale
SOC	System Organ Class
TEMA	treatment-emergent markedly abnormal
TRD	treatment resistant depression

1. INTRODUCTION

This statistical analysis plan (SAP) contains definitions of analysis sets, derived variables, and statistical methods for the study CNTO136MDD2001.

1.1. Trial Objectives

Primary Objective

To evaluate the efficacy of sirukumab as adjunctive treatment to antidepressant therapy (monoaminergic antidepressant) where sirukumab (administered as a 50mg subcutaneous (SC) injection at Day 1, Day 28 and Day 56 during the 12-week double-blind treatment period) is compared to adjunctive placebo based on the change from baseline to 12-week endpoint in depressive symptoms as measured by the total score on the Hamilton Depression Rating Scale ($HDRS_{17}$), in subjects diagnosed with Major Depressive Disorder (MDD) who have had a suboptimal response to the current standard oral antidepressant therapy and have a screening high sensitivity C-Reactive Protein (hsCRP) ≥0.300 mg/dL (International System of Units (SI) 3.00 mg/L).

Secondary Objectives

The secondary objectives are to include the evaluation of the impact of treatment with adjunctive sirukumab compared to adjunctive placebo, on: (1) overall safety and tolerability (2) anhedonia (as measured by the change in the Snaith Hamilton Pleasure Scale [SHAPS] from baseline to week 12) (3) global severity of symptoms of MDD (as measured by the change in the Clinical Global Impression - Severity [CGI-S] scale from baseline to Week 12) (4) remission ($HDRS_{17}$ total score ≤7) and response rates (≥ 50% improvement in $HDRS_{17}$ total score from baseline) at 12-week end point (5) subject-reported severity of symptoms of MDD (as measured by the change in the Patient Health Questionnaire [PHQ-9] from baseline to Week 1 and to Week 12) (6) fatigue (as measured by the change in the Functional Assessment of Chronic Illness Therapy-Fatigue [FACIT Fatigue] scale from baseline to Weeks 1, 4, 8 and 12) and (7) to evaluate the efficacy measured by change in the HDRS17 scores following adjunctive sirukumab compared to adjunctive placebo in the Treatment Resistant Depression (TRD) versus non-TRD subjects. (8) Evaluate pharmacokinetics and immunogenicity of sirukumab in subjects with MDD. (9) Evaluate the impact of treatment with adjunctive sirukumab in subjects with screening hsCRP≥3.00 mg/L and in those with screening hsCRP<3.00 mg/L based on changes from baseline to Week 12 in $HDRS_{17}$ total score. (10) Assess the efficacy of sirukumab compared to placebo as adjunctive therapy to an antidepressant in subjects with MDD with screening hsCRP levels <0.300 mg/dL versus those with screening hsCRP levels ≥ 0.300 mg/dL in improving response of depressive symptoms, defined as the proportion of subjects having a ≥50% improvement in HDRS17 total score from baseline to the end of Week 12. (11) Assess the efficacy of sirukumab compared to placebo as adjunctive therapy to an antidepressant in subjects with MDD with screening hsCRP levels <0.300 mg/dL versus those with screening hsCRP levels ≥ 0.300 mg/dL in achieving remission of depressive symptoms, defined as the proportion of subjects having a $HDRS_{17}$ total score ≤7 at the end of Week 12. (12) Evaluate whether hsCRP levels correlate with clinical efficacy as measured by the change from baseline to Week 12 in $HDRS_{17}$ total score.

Exploratory Objectives

(1) To explore the efficacy of adjunctive sirukumab compared to adjunctive placebo, based on the change from baseline to 12-week endpoint in the Inventory of Depressive Symptomatology – Clinician Rated 30 Item Scale (IDS-C30). (2) To explore immune-system related biomarkers (hsCRP, cytokines, chemokines), as well as other mood disorder related biomarkers (growth factors, hypothalamic–pituitary–adrenal (HPA) axis markers, metabolic markers) and to explore the potential correlation between these biomarkers and the clinical response, non-response or safety parameters of sirukumab. (3) To explore genetic and epigenetic variation that may be related to clinical response, non-response, or safety parameters of sirukumab (4) To explore the effect of sirukumab on the pharmacokinetics of monoaminergic antidepressants. (5) To explore the relationship between the severity of childhood trauma as assessed through the Childhood Trauma Questionnaire and the change in $HDRS_{17}$ scores following adjunctive sirukumab or placebo.

1.2. Trial Design

This is a multicenter, double-blind, placebo-controlled study in male and female subjects, 21 to 64 years of age inclusive, with major depressive disorder (MDD) who have had a suboptimal response to standard oral antidepressant therapy; have failed no more than 3 antidepressant treatments in the current major depressive episode.

The study will consist of 3 phases: a screening phase of up to 4 weeks, a 12-week double-blind treatment phase, and a 14-week posttreatment follow-up phase. The total duration of subject participation will be approximately 26 weeks.

1.3. Statistical Hypothesis for Trial Objective(s)

The primary hypothesis of this study is that sirukumab will determine significant improvement in depressive symptoms from baseline compared to placebo when administered as an adjunctive treatment to a monoaminergic antidepressant in the treatment of patients with MDD who have shown a suboptimal response to standard oral antidepressant therapy and have a screening hsCRP ≥ 0.300 mg/dL (SI 3.00 mg/L), demonstrated by improvement in depressive symptoms from baseline to 12-week endpoint in the Hamilton Depression Rating Scale (HDRS17) total score.

1.4. Sample Size Justification

Approximately 192 subjects will be enrolled in this study, with approximately 96 subjects randomized per treatment group.

The target study population is male and female subjects with MDD who have had a suboptimal response to standard oral antidepressant therapy; have failed no more than 3 antidepressant treatments in the current major depressive episode; approximately 142 subjects with screening hsCRP ≥0.300 mg/dL (SI 3.00 mg/L) and approximately 50 subjects with screening hsCRP<0.300 mg/dL (SI 3.00 mg/L) will be enrolled.

The expected final sample size of 142 for subjects with screening hsCRP ≥ 0.300 mg/dL (SI 3.00 mg/L) was determined based on the assumption of an effect size of at least 0.5 for the HDRS17 (mean change from baseline to week 12 endpoint between the sirukumab and placebo groups of 4 with SD=8). This is considered to be a clinically relevant difference in a population with suboptimal response to standard oral antidepressant therapy. Power is set at 90.0%, with a 1-sided alpha of 0.125 and a 12-week drop-out rate of 25%. It was also assumed that 10% of the randomized subjects would be excluded from the primary efficacy analysis due to having hsCRP ≥3.00 mg/L at screening but not at baseline.

1.5. Randomization and Blinding

Central randomization was implemented in this study. Subjects were randomly assigned to 1 of 2 treatment groups based on a computer-generated randomization schedule prepared before the study by or under the supervision of the sponsor. The randomization was balanced by using randomly permuted blocks and was stratified by country and 4 screening hsCRP levels: < 3.00 mg/L (0.300 mg/dL), ≥3.00 to <5.00 mg/L (≥0.300 to <0.500 mg/dL), ≥5.00 to <8.00 mg/L (≥0.500 to <0.800 mg/dL), and ≥8.00 mg/L (≥0.800 mg/dL). The interactive web response system (IWRS) assigns a unique treatment code, which dictates the treatment assignment and matching study drug kit for the subject. The requestor must use his or her own user identification and personal identification number when contacting the IWRS, and was then give the relevant subject details to uniquely identify the subject.

In general, randomization codes will be disclosed fully only if the study is completed and the clinical database is closed. At the interim analyses, the randomization codes and, if required, the translation of randomization codes into treatment and control groups will be disclosed to those authorized and only for those subjects included in the interim analysis.

2. GENERAL ANALYSIS DEFINITIONS

2.1. Treatment Groups

Unless specified, the treatment groups are labeled as ‘Placebo’ and ‘Sirukumab’.

2.2. Pooling Algorithm

Five counties are planned to enroll subjects: US, Canada, United Kingdom, Poland and Russia. To account for country variability, country will be used as a factor in the statistical models to analyze efficacy.

2.3. Analysis Phases

There are 2 analysis phases defined in this study: Double-blind and Follow-up Phases. Each analysis phase has its own analysis reference start date and end dates.

2.3.1. Analysis Phase Start and End Dates

Double-Blind Phase

The analysis reference start date of the DB analysis phase is the date of the first dose of DB medication. The analysis reference end date of the DB analysis phase is the maximum of date of the last visit in the DB treatment phase or date of early termination from DB phase. For randomized subjects who did not receive any medication in the DB phase, the reference start date is missing.

Follow-Up Phase

The analysis reference start date of the Follow-up (Post-treatment) analysis phase is the day after the end of DB. The analysis reference end date is the maximum of the last follow-up visit date and end of trial date.

2.3.2. Study Reference Start and End Dates

The overall reference start date for the study is defined as the date of the first dose of DB medication (the date is missing for screened subjects who did not receive a dose of DB medication). The overall reference end date for the study is maximum of the last follow-up visit date and end of trial date.

2.3.3. Study Day and Relative Days

Study day is calculated relative to the overall reference start date for the study. Relative day is calculated relative to the analysis reference start date of the analysis phase in which the data are captured.

Study day for an event on or after the start of the study is calculated as:

Event date-study start date +1.

Study day for an event prior to the start of the study is calculated as:

Event date-study start date

Relative day for an event on or after a reference start date is calculated as:

Event date – reference start date + 1.

Relative day for an event prior to a reference start date is calculated as:

Event date – reference start date.

2.4. Visit Windows

As subjects do not always adhere to the protocol visit schedule, the following rules are applied to assign actual visits to protocol visits. Listed below are the visit windows and the target days for each visit. The reference day is Study Day 1 (which is the date of the first DB medication in the double-blind phase).

The baseline value is defined as the last observation before receiving the first dose of study drug on Day 1.

If a subject has 2 or more actual visits in one visit window, the visit closest to the target day will be used as the protocol visit for that visit window. If 2 actual visits are equidistant from the target day within a visit window, the later visit is used. If a visit window has no scheduled visits but does have unscheduled visits, then the unscheduled visit closest to the scheduled visit will be used. If a visit window has both a scheduled and an unscheduled visit, then a scheduled visit will have a priority even an unscheduled visit is closer to the target day.

All assignments will be made in chronological order. Once a visit is assigned to a visit window, it will no longer be used for a later time point except for the end point. Listed below are the visit windows and the target days (if applicable) for each visit defined in the protocol (Table 1).

Table 1: Analysis Visit Windows			
Parameter	**Scheduled Visit Time (label on output)**	**Time Interval (Day)**	**Target Time Point (Day)**
$HDRS_{17}$	Baseline	1/predose	1
	Week 1 (DB)	2 - 17	7
	Week 4 (DB)	18- 42	28
	Week 8 (DB)	43 - 70	56
	Week 12 (DB)	71 - end of DB	84
CGI-S, $IDS\text{-}C_{30}$ and PHQ-9, SHAPS and FACIT Fatigue	Baseline	1/predose	1
	Week 1 (DB)	2 - 17	7
	Week 4 (DB)	18- 42	28
	Week 8 (DB)	43 - 70	56
	Week 12 (DB)	71 - end of DB	84
	Week16 (FU)	End of DB+1 - 133	112
	Week 22 (FU)	134 – end of FU	154
C-SSRS	Baseline	1/predose	1
	Week 1 (DB)	2 - 17	7
	Week 4 (DB)	18- 42	28
	Week 8 (DB)	43 - 70	56
	Week 12 (DB)	71 - end of DB	84
	Week16 (FU)	End of DB+1 - 133	112
	Week 22 (FU)	134 – end of FU	154
Vital signs (BP, HR, RR, temperature)	Baseline	1/predose	1
	Week 4 (DB)	2 - 56	28
	Week 12 (DB)	57 - end of DB	84
	Week16 (FU)	End of DB+1 - 133	112
	Week 22 (FU)	134 – end of FU	154
Body weight	Baseline	1/predose	1
	Week 12	2 – end of DB	84
	Week 22 (FU)	End of DB+1 – end of FU	154
Clinical laboratory assessments: chemistry, hematology	Baseline	1/predose	1
	Week 4	2 - 42	28
	Week 8	43 - 70	56
	Week 12	71 – end of DB	84
	Week 22	End of DB+1 – end of FU	154
Clinical laboratory assessments: urinalysis	Baseline	1/predose	1
	Week 8	2 - 70	56
	Week 12	71 – end of DB	84
	Week 22	End of DB+1 – end of FU	154

2.5. Analysis Sets

Subjects will be classified mainly into the following analysis sets: all randomized analysis set, intent-to-treat analysis sets (with different variations) and safety analysis sets.

2.5.1. All Randomized Analysis Set

This analysis set will include all subjects who were randomized (i.e., subjects who reported a randomization date, or were assigned a randomization number) regardless of whether or not the treatment was received. This analysis set will be used for summarizing the overall study completion/withdrawal information using planned treatment group.

2.5.2. Efficacy Analysis Sets

Three analyses sets will be defined for efficacy analyses: the modified intent-to-treat 1 (mITT1), the modified intent-to-treat 2 (mITT2) and the modified intent-to-treat 3 (mITT3) analyses sets.

The mITT1 analysis set is defined as all randomized subjects with hsCRP ≥ 3.00 mg/L at screening and baseline who receive at least 1 dose of study drug and have both the baseline and at least one postbaseline $HDRS_{17}$ total score measured within the double-blind treatment period.

The mITT2 analysis set is defined as all randomized subjects who receive at least 1 dose of study drug and have both the baseline and at least one postbaseline $HDRS_{17}$ total score measured within the double-blind treatment period.

The mITT3 analysis set is defined as all randomized subjects with hsCRP≥ 3.00 mg/L at screening who receive at least 1 dose of study drug and have both the baseline and at least one postbaseline $HDRS_{17}$ total score measured within the double-blind treatment period.

Subjects who receive an incorrect treatment will be analyzed under the planned treatment received.

2.5.3. Biomarker Analysis Set

Biomarker analyses will be performed on biomarker analysis set. It will include all randomized subjects who receive at least 1 dose of study drug in the double-blind phase and have at least 1 sample obtained.

2.5.4. Pharmacokinetic and Immunogenicity Analyses Sets

Pharmacokinetic analyses will be performed on pharmacokinetic analysis set. It will include all randomized subjects who receive at least 1 administration of sirukumab in the double-blind phase and have at least 1 sample obtained after sirukumab treatment.

Immunogenicity analyses will be performed on immunogenicity analysis set. It will include all randomized subjects who receive at least 1 administration of sirukumab and have appropriate samples for detection of antibodies to sirukumab (ie, subjects with at least 1 sample obtained after sirukumab treatment).

2.5.5. Safety Analysis Set

Safety analyses will be performed on the safety analysis set. It will include all randomized subjects who receive at least 1 dose of study drug in the double-blind phase. Analyses of change from baseline will include only subjects who have both baseline and at least 1 post-baseline observation during the double-blind phase.

Safety analyses for the follow up phase will be performed on the Safety (FU) analysis set. It will include all subjects who have at least 1 visit during the follow up phase.

Screen failures and randomized subjects who received no double-blind study medication will be excluded from the safety analysis set. Subjects who received an incorrect treatment will be analyzed under the planned treatment.

2.6. Definition of Subgroups

Descriptive statistics will be performed for the primary endpoint of $HDRS_{17}$ change from baseline stratified by the following sub-groups.

- Screening hsCRP stratification (<3.00 mg/L, ≥3.00 to <5.00 mg/L, ≥5.00 to <8.00 mg/L and ≥8.00 mg/L)
- Treatment Resistant Depression (TRD) vs. non-TRD subjects. (TRD is defined as having be treated with ≥ 2 trials of antidepressants of adequate dose and duration during the current episode.)

Other subgroups such as 5 types of maltreatment (physical abuse, physical neglect, sexual abuse, emotional abuse and emotional neglect) based on CTQ and baseline BMI status (≤25 (normal) vs. >25 (abnormal) or ≤30 (normal) vs. >30 (abnormal)) may also be explored.

2.7. Incomplete/Missing Dates for Adverse Events

Treatment-emergent adverse events (AEs) for DB phase are those events with an onset date on or after the start of double-blind study medication, and occurred on or before the end of the double-blind phase. A conservative approach will be used to handle the missing dates for adverse events.

The rules for estimating incomplete AE onset dates will be as follows:

(1) The missing day of the month will be estimated as follows: If the month and year are known and double-blind study medication started during that month then the estimated date is the start date of double-blind study medication. If the month and year are known and double-blind study medication started prior to that month then the estimated date is the 1st day of the month. If the month and year are known and double-blind study medication started after the month, then no estimation will be done, and the AE will not be considered as treatment emergent for the double-blind phase.

(2) If both the day and the month are missing: No estimation will be performed. However, these AEs will be considered treatment emergent for the double-blind phase and will be included in the double-blind treatment summaries, except for the calculation of duration of the AE. Attempts will be made to get at least the month for the adverse events.

For incomplete AE resolution dates, the rules are:

(1) The missing day of the month will be estimated as follows: If the month and year are known and the study medication was stopped before, or during that month, the estimated date is the last day of the month or the end of the double-blind phase, whichever is earlier. If the study medication stopped after that month then the estimated date is the last day of the month.

(2) If both the day and the month are missing: the estimated resolution date is the end of the double-blind phase.

3. SUBJECT INFORMATION

3.1. Demographics and Baseline Characteristics

Demographic and baseline characteristics (Table 2) and psychiatric history at baseline (Table 3) will be summarized by treatment group for the All Randomized analysis set. The continuous variables will be summarized using descriptive statistics (N, mean, standard deviation [SD], median, minimum, and maximum). The categorical variables will be summarized using a frequency distribution with the number and percentage of subjects in each category.

Table 2: Demographic Variables and Baseline Characteristics
Continuous Variables:
• Age (years) (informed consent date – date of birth + 1) / 365.25
• Baseline weight (kg)
• Baseline height (cm)
• Baseline BMI (kg/m2) calculated as Weight (kg)/[Height (m)]^2
• Baseline hsCRP level
Categorical Variables:
• Sex (male, female)
• Race[a] (White, Black or African American, Asian, American Indian or Alaskan native,
Native Hawaiian or Pacific islander, other)
• Ethnicity (Hispanic or Latino, not Hispanic or Latino)
• Baseline hsCRP (< 3.00 mg/L, ≥3.00 to <5.00 mg/L, ≥5.00 to <8.00 mg/L and ≥8.00 mg/L)

a. If multiple race categories are indicated, then Race is recorded as "Multiple".

Table 3: Psychiatric History at Baseline Variables
Continuous Variables:
• Baseline $HDRS_{17}$
• Baseline IDS-C30 total score
• Baseline CGI-S score
Categorical Variables:
• Baseline CGI-S score
• Antidepressant treatment history (as obtained in the MGH-ATRQ)

3.2. Disposition Information

The distribution of the number of subjects who are randomized, receive double-blind treatment, and complete the Week 12 will be presented by treatment group for the double-blind phase of the study. In addition, the distribution of trial termination reasons will be summarized. These summaries will include a total group. The All Randomized Subjects analysis set will be used to summarize this information. A subject will be considered to have completed the Week 12 treatment phase if he or she has completed assessments through Week 12 visit.

3.3. Extent of Exposure

Total duration of exposure during the DB phase is defined as days between the first and the last dose of DB study medication.

Descriptive statistics (N, mean, median, and range) of total duration of exposure will be presented for DB phase by treatment group in safety analysis set. A frequency distribution showing the number of doses received during the double-blind phase will also be provided.

3.4. Protocol Deviations

Deviations that occurred during the study will be listed.

3.5. Prior and Concomitant Medications

The number and percent of subjects who receive concomitant therapies will be summarized by planned treatment group and by the generic term of the medication for the safety analysis set. Concomitant medications will be summarized in DB and FU separately.

A similar summary will be presented for antidepressants received prior to study entry according to responses to the MGH-ATRQ for the Safety analysis set.

4. EFFICACY

Efficacy data will be provided for the interim efficacy analysis; however, they will not be provided for the first DMC Safety Review meeting, except for Item 3 (Suicide) of $HDRS_{17}$ and Item 18 (Suicide Ideation) of IDS-C30 which will be provided for DMC Safety Review meetings.

The efficacy variables in this study are listed in Table 4.

Table 4: Efficacy Variables

Scale	End point
Change in $HDRS_{17}$ total score from baseline to Week 12	Primary
Change in $HDRS_{17}$ total score from baseline to Weeks 1, 4, 8, 16 or 22	Secondary
Response rate (≥ 50% improvement in $HDRS_{17}$ total score from baseline) at Week 12	
Remission rates ($HDRS_{17}$ total score ≤7) at Week 12	
Change in CGI-S score from baseline to Weeks 1, 4, 8, 12, 16 and 22	
Change in SHAPS total score from baseline to Weeks 1, 4, 8, 12, 16 and 22	
Change in FACIT Fatigue total score from baseline to Weeks 1, 4, 8, 12, 16 and 22	
Change in IDS-C30 from baseline to Weeks 1, 4, 8, 12, 16 and 22	Exploratory
Change in PHQ-9 total score from baseline to Weeks 1, 4, 8, 12, 16 and 22	

4.1. Analysis Specifications

4.1.1. Level of Significance

Unless otherwise specified, all statistical comparisons will be carried out at a 1-sided significance level of 12.5% and confidence intervals will be presented at 2-sided confidence level of 75% (ie. equivalent to 1-sided confidence level of 87.5%).

4.1.2. Data Handling Rules

The last observation carried forward (LOCF) method will be applied to the $HDRS_{17}$ total score in the sensitivity analysis.

The LOCF approach will be used for imputing missing visit data in the ITT LOCF sensitivity efficacy analyses. The LOCF calculation will be performed for all scheduled post-baseline time windows in the double-blind phase.

4.1.3. Imputation Methods for Missing Items

Imputation of missing individual item scores will apply only to the $HDRS_{17}$, as described in the next section. For all other scales where multiple items are summed to create a total, if any item of the scale is missing on one visit, the total score for that scale at that visit will be left blank.

4.2. Primary Efficacy Endpoint

4.2.1. Definition

The $HDRS_{17}$ is a clinician-administered rating scale designed to assess the severity of symptoms in subjects diagnosed with depression with a score range of 0 to 52. Each of the 17 items is rated by the clinician on either a 3- or a 5-point scale. The $HDRS_{17}$ consists of 17 items that cover all of the core depressive symptoms (depressed mood, feeling of guilt, suicide, insomnia early, insomnia middle, insomnia late, work and activities, retardation [psychomotor], agitation, anxiety [psychological], anxiety somatic, somatic symptoms general, genital symptoms, hypochondriasis, loss of weight and insight). A total score (0 to 52) is calculated by adding the scores of all 17 items. For each item as well as the total score, a higher score represents a more severe condition. If 2 or more items are missing, no imputation will be performed and the total score will be left missing. Otherwise, the total score will be calculated as the sum of the items present multiplied by the ratio of the maximum possible number of items (i.e., 17) to the number of items present.

4.2.2. Analysis Methods

The primary efficacy analyses will be conducted using mITT1 analysis set. The primary comparison will be between sirukumab and placebo using the primary efficacy endpoint, i.e., change in $HDRS_{17}$ total score from baseline to Week 12.

A mixed effects model using repeated measures (MMRM) with time, treatment, country, screening hsCRP stratification and time-by-treatment interaction as factors, baseline $HDRS_{17}$ total score as a continuous covariate, and a random subject effect, will be conducted. An unstructured variance-covariance matrix will be used (if there are convergence issues with the unstructured variance-covariance matrix, another suitable matrix structure will be chosen based on the LR test or the AIC). The contrast on Week 12 changes will be of primary interest, and tested at one-sided alpha level of 0.125. Descriptive statistics for values and changes from baseline will be provided for $HDRS_{17}$ including individual items and total score, at each time point of the double-blind treatment phase. Time profile of $HDRS_{17}$ total score for individual subjects will be represented graphically over time by treatment group.

Sensitivity analyses on the primary endpoint will also be conducted. Sensitivity analyses will be using the aforementioned MMRM using mITT3 and mITT2 analysis sets.

The primary endpoint with the aforementioned MMRM analyses may also be explored in subgroups such as 5 types of maltreatment based on CTQ (physical abuse, physical neglect, sexual abuse, emotional abuse and emotional neglect) and baseline BMI status ($\leq$25 (normal) vs. >25 (abnormal) or $\leq$30 (normal) vs. >30 (abnormal)).

In addition, the primary endpoint using similar MMRM analyses with additional covariates such as CTQ total score or others may also be explored.

Secondary Efficacy Endpoints
Analyses of secondary efficacy endpoints will be mainly based on the mITT2 analysis set. Analyses on the secondary efficacy endpoints may also be conducted using the aforementioned MMRM using mITT1 and mITT3 analysis sets.

4.2.3. $HDRS_{17}$ Total Score

Changes in $HDRS_{17}$ from baseline to Weeks 1, 4, and 8 will be analyzed using the same mixed-effects model with repeated measures (MMRM) as for the primary endpoint in all 3 mITT analysis datasets.

Summary statistics on $HDRS_{17}$ (changes from baseline) over time by treatment group will be provided for the following subgroups:
(1) Screening hsCRP stratification;

(2) TRD status (TRD vs. non-TRD);

Other subgroups such as 5 types of maltreatment (physical abuse, physical neglect, sexual abuse, emotional abuse and emotional neglect) based on CTQ and baseline BMI status (≤25 (normal) vs. >25 (abnormal) or ≤30 (normal) vs. >30 (abnormal)) may also be explored.

In addition, mean time profile plots of $HDRS_{17}$ over time by treatment group will be provided.

4.2.4. Responders and Remitters

Subjects who have $HDRS_{17}$ score ≤ 7are considered as remitters while those who have ≥ 50% improvement in $HDRS_{17}$ total score from baseline at week 12 are considered responders. The response rate or remission rate at Week 12 in sirukumab will be compared with that in placebo using a logistic regression model including baseline $HDRS_{17}$ score, screening hsCRP stratification, country and treatment.

Response rate or remission rate at Week 12 will be provided by treatment group for the following subgroups:
(1) Screening hsCRP stratification;

(2) TRD status (TRD vs. non-TRD).

4.2.5. Clinical Global Impression-Severity Scale (CGI-S)

The CGI-S is a clinician -rated scale that assesses the severity of mental illness with scores as follows:

1: normal, not at all ill; 2: borderline mentally ill; 3: mildly ill; 4: moderately ill; 5: markedly ill; 6: severely ill; 7: among the most extremely ill patients.

A higher score implies a more severe condition. Change from baseline in CGI-S score will be calculated.

Descriptive statistics (N, mean, median and range) of the numerical scores and as well changes from baseline will be provided by treatment group.

A frequency distribution of severity scores may be provided by treatment group from Week 1 to Week 12.

In addition, analyses on changes from baseline in CGI-S score may also be conducted using the similar aforementioned MMRM for the primary analyses.

4.2.6. Patient Health Questionnaire (PHQ-9)

The PHQ-9 is used as a subject-reported measure of depressive symptomatology. The PHQ-9 is a 9-item scale, where each item is rated on a 4-point scale (0=Not at all, 1=Several Days, 2=More than half the days, and 3=Nearly every day), with a total score range of 0 to 27. The recall period is 2 weeks.

Descriptive statistics (N, mean, median and range) of total score and as well changes from baseline will be provided by treatment group.

In addition, analyses on changes from baseline in PHQ-9 total score may also be conducted using the similar aforementioned MMRM for the primary analyses.

4.2.7. Snaith–Hamilton Pleasure Scale (SHAPS)

Anhedonia, the inability to experience pleasure, is a core symptom of depression. The Snaith–Hamilton Pleasure Scale (SHAPS) is a short, 14-item instrument to measure anhedonia, which has been shown to be valid and reliable in normal and clinical samples. Each of the 14 items has a set of four response categories: Definitely Agree, Agree, Disagree, and Definitely Disagree.

There will be two definitions for total score. The (protocol) definition 1: Definitely Agree (=1), Agree (= 2), Disagree (= 3), and Definitely Disagree (= 4). The definition 2: Definitely Agree (= 0), Agree (= 0), Disagree (= 1), and Definitely Disagree (= 1). A SHAPS total score will be calculated as the sum of the 14 item scores. A higher total score indicates higher levels of state anhedonia.

Descriptive statistics (N, mean, median and range) of total score (for both definitions) and as well changes from baseline will be provided by treatment group.

4.2.8. Functional Assessment of Chronic Illness Therapy-Fatigue (FACIT Fatigue)

The FACIT-Fatigue is a questionnaire that assesses self-reported tiredness, weakness, and difficulty conducting usual activities due to fatigue. The subscale consists 13-item instrument to measure fatigue. Each of the 13 items has a set of five response categories: Not at all (=0), A little bit (=1), Somewhat (=2), Quite a bit (=3) and Very much (=4). A total FACIT-Fatigue subscale score will be calculated as the sum of the 13 item scores (reserved scores [4 – score] for all except for 2 items: "I have energy" and "I am able to do my usual activities"), and ranges from 0 to 52, with a higher score indicating less fatigue.

Descriptive statistics (N, mean, median and range) of total score and as well changes from baseline will be provided by treatment group.

4.3. Exploratory Efficacy Endpoint

4.3.1. Inventory of Depressive Symptomatology – Clinician Rated 30 (IDS-C30)

The clinician-rated IDS is a 30-item, depression-specific symptom severity rating scale. The IDS is designed to measure the specific signs and symptoms of depression, including melancholic, atypical and anxious features. Scores range from 0 to 84 with higher scores representing greater severity of depressive symptoms. Only one of items 11 or 12 and of items 13 or 14 need to be rated. If both item 11 and 12 (or 13 and 14) are completed by accident, the highest of the two item scores will be used for calculating the total score.

Change from baseline in IDS-C30 score will be calculated. Descriptive statistics (N, mean, median and range) of the numerical scores and as well changes from baseline will be provided by treatment group.

In addition, analyses on changes from baseline in IDS-C30 total score may also be conducted using the similar aforementioned MMRM for the primary analyses.

4.3.2. Childhood Trauma Questionnaire (CTQ)

The self-report includes a 28-item test that measures 5 types of maltreatment – emotional, physical, and sexual abuse, and emotional and physical neglect. A 5-point Likert scale is used for the responses which range from Never True to Very Often True. A total CTQ score will be calculated as the sum of the 28 item scores (reserved scores for items 2, 5, 7, 10, 13, 16, 19, 22, 26 and 28), and ranges from 28 to 140, with a higher score indicating a more server childhood trauma.

Physical abuse is 'yes' if sum of CTQ items 9, 11, 12, 15, 17 is 8 or larger. Physical neglect is 'yes' if sum of CTQ items 1, 2 (R) (where R stands for reserving score), 4, 6, 26 (R) is 8 or larger. Sexual abuse is 'yes' if sum of CTQ items 20, 21, 23, 24, 27 is 8 or larger. Emotional abuse is 'yes' if sum of CTQ items 3, 8, 14, 18, 25 is 10 or larger. Emotional neglect is 'yes' if sum of CTQ items 5 (R), 7 (R), 13 (R), 19 (R), 28 (R) is 15 or larger.

Descriptive statistics (N, mean, standard deviation, median and range) of the numerical total scores may be provided by treatment. A frequency distribution of CTQ will be provided by treatment group. Frequency distributions of physical abuse, physical neglect, sexual abuse, emotional abuse and emotional neglect will also be provided by treatment group.

4.3.3. MINI International Neuropsychiatric Inventory (MINI)

The Mini-International Neuropsychiatric Interview (M.I.N.I.) is a short structured diagnostic interview for psychiatric disorders.

Frequency distributions of diagnostic psychiatric disorders will be provided by treatment group.

5. BIOMARKERS

5.1. Biomarker Variables

Biomarkers consist of hsCRP, ACTH, Adiponectin, serum BDNF, IGF1, IL6R, Leptin, IL-6, IL-1b, IL-10, TNFa, saliva Cortisol, serum Cortisol, Kynurenine pathway metabolites, and a total of 92 SNPs related to above HPA, metabolic, and inflammatory biomarkers.

5.2. Analysis Methods

5.2.1. hsCRP as a predictor of response

HDRS17 changes and percent changes from screening on week 12 (Δ_{w12}^{HDRS17}) will be modeled as a function of screening hsCRP including BMI, sex, and age as covariates for subjects treated with sirukumab. hsCRP is determined to have significant predictive power if it models a significant amount of variance in Δ_{w12}^{HDRS17} from screening at $p <= 0.1$ Scatter plots of Δ_{w12}^{HDRS17} versus hsCRP will be created and annotated with regression modeling results including regression coefficients for model variables and corresponding significance levels and coefficients of determination R^2 and $R^2_{partial}$ and $R^2_{reduced}$ for the reduced model. The reduced model is the full model with the variable of interest (hsCRP) omitted.

Logistic models will be used to gauge the potential of screening hsCRP as a predictor of remitters (week 12 HDRS17 <= 7) and responders with (50% improvement in HDRS17 on week 12). In both models, covariates will include BMI, sex, and age. Screening HDRS17 levels will also be used as a covariate in the model predicting remission. ROC curves will be generated for each model in addition to AUC and model accuracies.

To demonstrate specificity of hsCRP predictive power, modeling is also carried out on placebo treated subjects and the outcomes are compared to those for subjects treated with sirukumab.

In order to enable comparison of hsCRP performance with that of other biomarkers which were measured at baseline, the above models will be repeated substituting screening visit with baseline.

5.2.2. Alternate predictors of response

5.2.2.1 Tier A biomarker panel

The Tier A biomarkers panel consists of inflammation-related analytes: IL-6, IL6R, IL-10, IL-1b, and TNFa. Kynurenine pathway metabolites kynurenine, tryptophan, 3-OH kynurenine, kynurenic acid, and quinolinic acid are also included in Tier A as this pathway has been shown to drive inflammatory processes.

Δ_{w12}^{HDRS17} will be modeled as a function of each analyte in the high-priority panel measured at baseline with BMI, sex, and age as covariates. Analyte levels that model a significant amount of variance ($p \leq 0.1$) in Δ_{w12}^{HDRS17} will be compared against performance of the baseline hsCRP model from Section 5.2.1 Model performance will be compared using Akaike Information Criterion (AIC) or non-parametric equivalent. Bonferroni Family Wise Error Rates (FWER) will be reported in addition to uncorrected p values. Models with FWER ≤ 0.1 and better performance than the hsCRP model are identified as alternate candidate predictive biomarkers.

Logistic models for identified alternate candidate biomarkers, if any, will be generated for predicting remission and response and the results will be compared to the corresponding outcome of logistic models for baseline hsCRP in Section 5.2.1.

Multivariate models using supervised machine learning will also be used to predict Δ_{w12}^{HDRS17}, response, and remission using hsCRP and Tier A analytes. 80% of the samples will be used for training with 20% set aside for model testing. ROC curves will be generated for each of model in addition to AUC and model accuracies.

5.2.2.2 Full biomarker panel

If analysis of high-priority analytes fails to identify alternate candidate predictive biomarkers, exploratory univariate modeling of Δ_{w12}^{HDRS17} will be carried out with each of the remaining baseline biomarkers. Modeling outcomes will be tabulated including uncorrected p-values as well as p-values adjusted by controlling the Benjamini & Hochberg False Discovery Rate (BH FDR). Scatter plots for regression models with uncorrected p-values ≤ 0.05 will be created and annotated with coefficients of determination (R^2 and $R^2_{partial}$) and $R^2_{reduced}$ for the reduced model.

Multivariate models using supervised machine learning will also be used to predict Δ_{w12}^{HDRS17}, response, and remission using all analytes that reach significance predicting Δ_{w12}^{HDRS17} at uncorrected $p \leq 0.1$ in the univariate models. 80% of the samples will be used for training with 20% set aside for model testing. ROC curves will be generated for each of model in addition to AUC and model accuracies. Results from multivariate modeling will also be reported using all analytes regardless of univariate significance.

6. PHARMACOKINETICS AND IMMUNOGENICITY

6.1. Pharmacokinetic Analyses

Serum sirukumab concentrations will be summarized over time. Descriptive statistics, including arithmetic mean, standard deviation (SD), median, interquartile range, minimum, and maximum will be calculated at each sampling time point. Monoaminergic antidepressants concentration data before and after sirukumab treatment, will also be summarized.

Data will be listed for all subjects with available serum concentrations.

All concentrations below the lowest quantifiable concentration or missing data will be labeled as such in the concentration database. All subjects and samples excluded from the analysis will be clearly documented in the study report.

6.2. Immunogenicity Analyses

The incidence of antibodies to sirukumab will be summarized for all subjects who received at least 1 administration of sirukumab and have appropriate samples for detection of antibodies to sirukumab (ie, subjects with at least 1 sample obtained after sirukumab treatment).

7. SAFETY

7.1. Adverse Events

Adverse events (AEs) are coded using the MedDRA dictionary (version 17.0 or above). Treatment-emergent adverse events (TEAEs) that occurred in the double-blind phase will be summarized by system organ class, preferred term, and treatment group. Adverse events that occurred in the follow up phase will be summarized separately.

Adverse events in different phases of the study are defined as below:

- Treatment-emergent adverse events during the double-blind phase are defined as AEs with an onset during the double-blind phase. In other words, treatment-emergent AE during the double-blind phase should satisfy the condition: Double-blind start date/time ≤ AE onset date/time ≤ double-blind end date. If onset time is missing and AE onset date/time is the same as the double-blind start date/time, the AE is defined to be treatment emergent in the double-blind phase.

- Adverse events during the follow up phase are defined as AEs with an onset during the follow up phase. AEs during the follow up phase should satisfy the condition: (follow up phase start date ≤AE onset date/time and AE onset date ≤ follow up phase end date).

A TEAE is an event that is new in onset or increased in severity following treatment initiation. An event that starts prior to, and ends after the initiation of study medication will be considered treatment-emergent only if the severity increases after the start of medication. Adverse events will not be considered treatment-emergent if they occur or increase in severity during the follow up phase. Adverse events occurring during the follow up phase will be summarized separately. In addition, AEs will be summarized by severity and relationship to study drug using the preferred term. For the summaries of AEs by severity/relationship to study drug, the observation with the most severe occurrence/closest relationship to study drug will be chosen if there is more than one incident of an adverse event reported during the analysis phase by the subject.

Serious AEs (SAEs) and AEs that lead to study discontinuation will be summarized separately by treatment group, system organ class, and preferred term. Data listings will also be generated for deaths, other SAEs, and discontinuations due to AEs.

Adverse Events of Special Interest

Clinically relevant TEAEs of special interest will be examined in malignancies, active TB, hepatobiliary abnormalities (as defined in the protocol), and gastrointestinal perforations.

7.2. Clinical Laboratory Tests

Descriptive statistics (N, mean, median and range) for values and changes from baseline will be provided for clinical laboratory tests (hsCRP, hematology, chemistry, lipid panel [fasting] and urinalysis) at each scheduled time point in the double-blind and follow up phases

In addition, shift table from baseline to post-baseline values with respect to the normal range will be provided.

7.3. Vital Signs, Weight and BMI

Descriptive statistics for values and changes from baseline at each scheduled time-point during the double-blind and follow up phases will be presented for Systolic Blood Pressure (mmHg), Diastolic Blood Pressure (mmHg), Pulse Rate (beats per minute), Weight (kg), and BMI. In addition, descriptive statistics of pulse rate and blood pressure (systolic and diastolic) values and changes from predose will be provided for each timepoint.

The proportion of subjects who have a treatment-emergent abnormality, as defined in Table 5 below, will be presented for the double-blind phase. The double-blind baseline will be used to determine abnormal values. A listing of subjects meeting any of the criteria will also be provided for the double-blind phase.

Table 5: Treatment-Emergent Abnormality Categories for Vital Signs

Vital Parameter	Post-baseline value outside of normal limit if:	
	Abnormally low	Abnormally high
Pulse (bpm)	A decrease from baseline of ≥ 15 to a value ≤ 50	An increase from baseline of ≥ 15 to a value ≥ 100
Systolic BP (mmHg)	A decrease from baseline of ≥ 20 to a value ≤ 90	An increase from baseline of ≥ 20 to a value ≥ 180
Diastolic BP (mmHg)	A decrease from baseline of ≥ 15 to a value ≤ 50	An increase from baseline of ≥ 15 to a value ≥ 105

BP = blood pressure

7.4. Other Safety Parameters

7.4.1. Columbia Suicide Severity Rating Scale

The C-SSRS is a low-burden measure of the spectrum of suicidal ideation and behavior. It is a semi structured clinician-administered questionnaire designed to solicit the occurrence, severity, and frequency of suicide-related ideation and behaviors during the assessment period.

As noted in the August 2012 draft guidance titled "Suicidal Ideation and Behavior: Prospective Assessment of Occurrence in Clinical Trials",[3] the FDA has outlined 11 categories of interest, obtained by the C-SSRS instrument. Some data collected on the C-

SSRS are not used for analysis, but are used for individual clinical management and safety monitoring (e.g., suicidal behavior lethality and suicidal ideation intensity). These data are not included in the proposed analysis tables.

The summaries of the C-SSRS outcomes will be based on the safety analysis set subjects who have at least 1 post-baseline C-SSRS measurement and a pre-treatment C-SSRS assessment (lifetime assessment at screening). The C-SSRS outcomes will be summarized in double-blind phase.

The following outcomes are C-SSRS categories and have binary responses (yes/no). The categories have been re-ordered from the actual scale to facilitate the definitions of the composite and comparative endpoints, and to enable clarity in the presentation of the results.

Categories

Category 1 – Wish to be Dead

Category 2 – Non-specific Active Suicidal Thoughts

Category 3 – Active Suicidal Ideation with Any Methods (Not Plan) without Intent to Act

Category 4 – Active Suicidal Ideation with Some Intent to Act, without Specific Plan

Category 5 – Active Suicidal Ideation with Specific Plan and Intent

Category 6 – Preparatory Acts or Behavior

Category 7 – Aborted Attempt

Category 8 – Interrupted Attempt

Category 9 – Actual Attempt (non-fatal)

Category 10 – Completed Suicide

Self-injurious behavior without suicidal intent is also a C-SSRS outcome (although not suicide-related) and has a binary response (yes/no).

The following outcome is a numerical score derived from the C-SSRS categories. The score is created at each assessment for each patient and is used for determining treatment emergence.

- Suicidal Ideation Score: The maximum suicidal ideation category (1-5 on the C-SSRS) present at the assessment. Assign a score of 0 if no ideation is present.

Composite endpoints based on the above categories are defined below.

- Suicidal ideation: A "yes" answer at any time during treatment to any one of the five suicidal ideation questions (Categories 1-5) on the C-SSRS.

- Suicidal behavior: A "yes" answer at any time during treatment to any one of the five suicidal behavior questions (Categories 6-10) on the C-SSRS.

Comparative endpoints of interest are defined below. "Treatment emergence" is used for outcomes that include events that first emerge or worsen. "Emergence" is used for outcomes that include events that first emerge.

- **Treatment-emergent suicidal ideation** compared to all prior history:

An increase in the maximum suicidal ideation score during treatment from the maximum suicidal ideation score prior to treatment.

- **Emergence of serious suicidal ideation** compared to all prior history:

An increase in the maximum suicidal ideation score to 4 or 5 during treatment from no suicidal ideation (scores of 0) prior to treatment.

- **Treatment-emergent serious suicidal ideation** compared to all prior history:

An increase in the maximum suicidal ideation score to 4 or 5 on the C-SSRS during treatment from not having serious suicidal ideation (scores of 0-3) prior to treatment.

- **Emergence of suicidal behavior** compared to all prior history:

The occurrence of suicidal behavior (any of the Categories 6-10) during treatment from not having suicidal behavior (any Categories 6-10) prior to treatment.

A summary of pre-treatment suicidal ideation and suicidal behavior will be presented in terms of percent of subjects with lifetime suicidal ideation, lifetime suicidal behavior, lifetime self-injurious behavior without suicidal intent.

Improvement in suicidal ideation compared to all prior history:

The occurrence of a suicidal ideation score (any of the Categories 1-5) during treatment lower than the suicidal ideation score (Categories 1-5) prior to treatment. Subjects with a

score of 0 prior to treatment (no ideation is present) will not be classified as having improved even if they did not experience any ideation during the study, ie, suicidal ideation score remains 0 post treatment.

The number and percent of subjects with treatment-emergent suicidal ideation, treatment emergent serious suicidal ideation, emergence of serious suicidal ideation and improvement in suicidal ideation and emergence of suicidal behavior in the double-blind phase will be summarized by treatment group in frequency and shift tables.

A bar chart that will display in a visual graph the improvements in suicidal ideation score by treatment sequence over time may be provided.

Outcomes that can be used for clinical management and safety monitoring are described below and will be presented in data listings.

8. REFERENCES

1. Mathew, S.J. et al., Riluzole for relapse prevention following intravenous ketamine in treatment-resistant depression: a pilot randomized, placebo-controlled continuation trial, International Journal of Neuropsychopharmacology, **13**, pp. 71-82, 2010.

2. International Conference on Harmonization (ICH) Tripartite Guideline E14. The clinical evaluation of QT/QTc interval prolongation and proarrhythmic for non-antiarrhythmic drugs. London 2005

3. FDA Guidance for Industry: Suicidal Ideation and Behavior: Prospective Assessment of Occurrence in Clinical Trials. August 2012