Official Title: A Phase 2, multicenter, randomized, double-blind, placebo-controlled, study to evaluate the efficacy and safety of adjunctive pimavanserin in major depressive disorder

NCT Number: NCT03018340

ACADIA®
Pharmaceuticals

STATISTICAL ANALYSIS PLAN

Protocol No.:	ACP-103-042
Protocol Title:	A Phase 2, Multicenter, Randomized, Double-blind, Placebo-controlled, Study to Evaluate the Efficacy and Safety of Adjunctive Pimavanserin in Major Depressive Disorder
Drug:	Pimavanserin
Sponsor:	ACADIA Pharmaceuticals Inc.
Version No. and Date	Version 1.0, 21 August 2018

SIGNATURE/APPROVAL PAGE

AUTHORS

[redacted] 10 SEP 2018
Date

Senior Director, Biostatistics
ACADIA Pharmaceuticals Inc.

APPROVERS

[redacted] 12 Sep 2018
Date

Senior Vice President, Clinical Development & CMO
ACADIA Pharmaceuticals Inc.

[redacted] 10 September 2018
Date

Executive Director, Clinical Research
ACADIA Pharmaceuticals Inc.

[redacted] 10 SEP 2018
Date

Vice President, Biometrics
ACADIA Pharmaceuticals Inc.

[redacted] 10 SEP 2018
Date

Executive Director, Biostatistics & SAS Programming
ACADIA Pharmaceuticals Inc.

TABLE OF CONTENTS

LIST OF TABLES

ABBREVIATIONS

AE	Adverse event
AIMS	Abnormal Involuntary Movement Scale
ANCOVA	analysis of covariance
ATC	Anatomical/Therapeutic/Chemical
BARS	Barnes Akathisia Rating Scale
BIS-11	Barratt Impulsiveness Scale
BMI	body mass index
CI	Confidence Interval
CGI-I	Clinical Global Impressions – Global Improvement
CGI-S	Clinical Global Impressions – Severity of Illness
C-SSRS	Columbia Suicide Severity Rating Scale
DMC	Data Monitoring Committee
DSM-5	Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition
eCRF	electronic case report form
ECG	electrocardiogram
FAS	Full Analysis Set
HAMD-17	Hamilton Depression Scale (17 Items)
IRT	Interactive Response Technology
KSS	Karolinska Sleepiness Scale
LOCF	last observation carried forward
MADRS	Montgomery Asberg Depression Scale
MDD	Major Depressive Disorder
MDE	Major Depressive Episode
MedDRA	Medical Dictionary for Regulatory Activities
MGH-ATRQ	Massachusetts General Hospital Antidepressant Treatment Questionnaire
MGH-SFI	Massachusetts General Hospital Sexual Functioning Index
MMRM	mixed model for repeated measures
OC	observed cases
PCI	potentially clinically important
PD	Pharmacodynamic(s)
PK	Pharmacokinetic(s)
QTcB	QT Interval Corrected for Heart Rate using Bazett's Formula
QTcF	QT Interval Corrected for Heart Rate using Fridericia's Formula

SAE	serious adverse event
SAP	statistical analysis plan
SCID-5-CT	Structured Clinical Interview for DSM-5, Clinical Trials Version
SD	Standard Deviation
SDS	Sheehan Disability Scale
SE	Standard Error
SF-12	12-Item Short Form Health Survey
SIS	Sheehan Irritability Scale
SNRI	serotonin-norepinephrine reuptake inhibitors
SOC	system organ class
SPCD	Sequential Parallel Comparison Design
SSRI	selective serotonin reuptake inhibitors
TEAE	treatment-emergent adverse event

1. INTRODUCTION

This statistical analysis plan (SAP) provides a technical and detailed elaboration of the statistical analyses of efficacy and safety data as described in Protocol Amendment 1 dated 26 April 2017. Specifications for tables, figures, and listings are contained in a separate document.

2. OBJECTIVES

2.1 Primary Objective

The primary objective of this study is to assess the efficacy of pimavanserin compared to placebo when given adjunctively to a selective serotonin reuptake inhibitor (SSRI)/serotonin-norepinephrine reuptake inhibitor (SNRI) antidepressant as treatment of patients with Major Depressive Disorder (MDD) and an inadequate response to antidepressant therapy.

2.2 Secondary Objectives

Secondary objectives of this study are to evaluate the efficacy of pimavanserin compared with placebo for the following:

- patient disability
- clinician's global assessment of treatment benefit
- quality of life
- drug attitude
- sleep
- sexual functioning
- impulsivity
- irritability

2.3 Other Secondary Objectives

- To assess the safety and tolerability of adjunctive pimavanserin compared to placebo
- To characterize the PK of pimavanserin administered as adjunct to an SSRI/SNRI antidepressant in MDD patients
- To assess the PK/PD using measures of safety, efficacy, and sleep parameters

3. STUDY DESIGN

3.1 General Study Design

This study is a multicenter, randomized, double-blind, placebo-controlled, 2-stage Sequential Parallel Comparison Design (SPCD) study in patients with MDD and an inadequate response to antidepressant therapy with concurrent SSRI/SNRI. MDD and Major Depressive Episode (MDE) are defined according to the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5), as confirmed by the Structured Clinical Interview for DSM-5, Clinical Trials Version (SCID-5-CT). Inadequate treatment response will be determined through the administration of the Massachusetts General Hospital Antidepressant Treatment Questionnaire (MGH ATRQ).

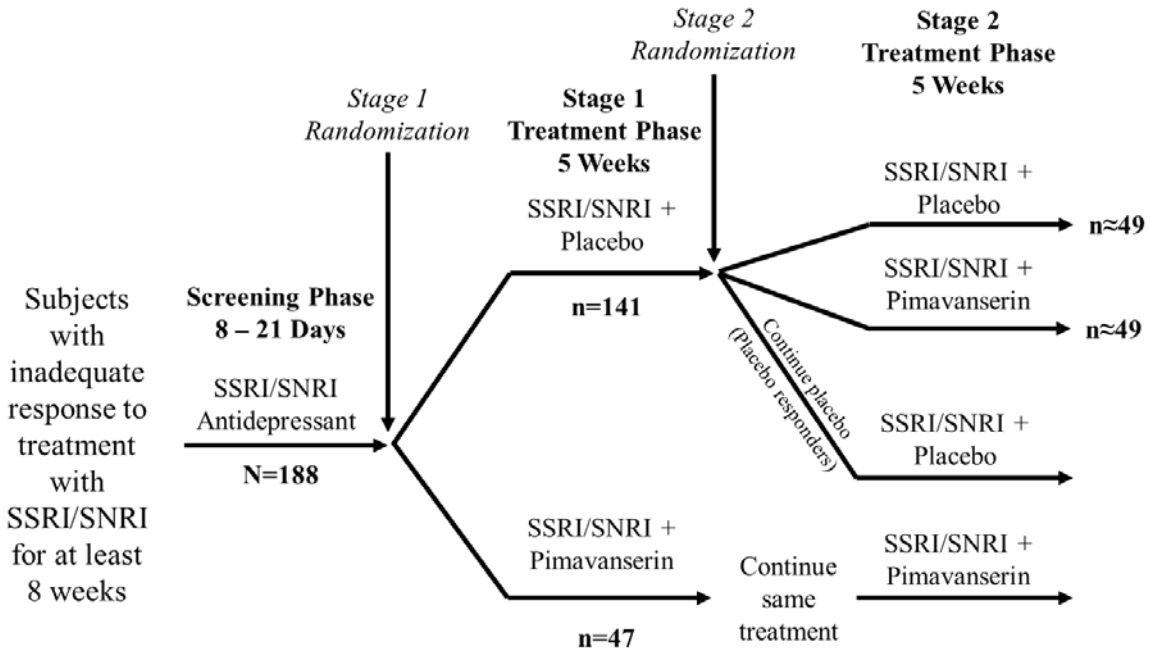
The study drug will be pimavanserin 34 mg (provided as 2×17 mg tablets) and matching placebo (2×placebo tablets), administered orally as a single dose once daily (QD).

The study duration is approximately 17 weeks consisting of a Screening Period of 8-21 days, a 10-week Treatment Period divided into 2 stages of equal 5-week durations, and a Safety Follow-up period of approximately 30 days. At the beginning of Stage 1 (Week 0), eligible subjects will be randomly assigned in a 1:3 ratio to pimavanserin or placebo, respectively. At the end of Stage 1 (Study Week 5), subjects initially randomized to placebo and who meet the non-responder criteria (i.e., HAMD-17 total score at Week 5 >14 and percent-reduction from the Study Week 0 (Baseline) HAMD-17 total score of <50%) will be randomly assigned in a 1:1 ratio to pimavanserin or placebo. Subjects who do not meet the criteria for re-randomization will continue with the assigned treatment from Stage 1 until the end of Stage 2. Subjects initially randomized to pimavanserin in Stage 1 will continue to receive pimavanserin through the end of Stage 2.

The study is a multicenter study with up to 30 study sites participating. Approximately 188 adult male and female patients with MDD are expected to participate in this study.

The study design is summarized in Figure 1.

Figure 1 Schematic of Study Design



3.2 Schedule of Assessments

Table 1 Schedule of Assessments

Visit		Double-blind Treatment Period											~30-Day Follow-Up[a]
	Screening 1	Baseline 2	3	4	5	6	7	8	9	10	11	12	
	(Day -21 to -8)	Week 0	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Week 7	Week 8	Week 9	Week 10/ET	
Visit window (days)			±3	±3	±3	±3	±3	±3	±3	±3	±3	±3	+7
Informed consent and if applicable, privacy forms	X												
Inclusion/exclusion criteria	X	X											
Medical history, medication history, and demographics	X												
Psychiatric history	X												
MGH ATRQ	X												
SAFER remote interview	X												
SCID-5-CT	X												
Physical examination	X	X					X					X	
Vital signs	X	X	X	X	X	X	X	X	X	X	X	X	
Height and weight[b]	X	X	X	X	X	X	X	X	X	X	X	X	
12-lead ECG[c]	X	X	X				X	X				X	
Clinical laboratory tests[d,e]	X	X					X					X	
Pregnancy test[f]	X	X					X					X	
PK blood draws[g]	X	X	X		X		X	X		X		X	
Urine toxicity screen	X	X					X						
MADRS	X	X											
HAMD-17	X	X	X	X	X	X	X	X	X	X	X	X	
CGI-S	X	X	X	X	X	X	X	X	X	X	X	X	
CGI-I			X	X	X	X	X	X	X	X	X	X	
SF-12		X					X					X	
SDS		X	X	X	X	X	X	X	X	X	X	X	

		Double-blind Treatment Period											~30-Day Follow-Up[a]
Visit	**Screening 1**	**Baseline 2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	
	(Day -21 to -8)	**Week 0**	**Week 1**	**Week 2**	**Week 3**	**Week 4**	**Week 5**	**Week 6**	**Week 7**	**Week 8**	**Week 9**	**Week 10/ET**	
DAI-10		X					X					X	
MGH-SFI		X					X					X	
KSS		X	X	X	X	X	X	X	X	X	X	X	
BIS-11		X	X		X		X	X		X		X	
SIS		X	X		X		X	X		X		X	
C-SSRS[h]	X	X	X	X	X	X	X	X	X	X	X	X	
AIMS		X	X				X	X				X	
BARS		X	X				X	X				X	
SAS		X	X				X	X				X	
Concomitant Medications	X	X	X	X	X	X	X	X	X	X	X	X	X
Assessment of adverse events	X	X	X	X	X	X	X	X	X	X	X	X	X
Dispense study drug		X	X	X	X	X	X	X	X	X	X		
Study drug accountability			X	X	X	X	X	X	X	X	X	X	

Abbreviations: AIMS = Abnormal Involuntary Movement Scale; BARS = Barnes Akathisia Rating Scale; BIS-11 = Barratt Impulsiveness Scale; CGI-I = Clinical Global Impression-Improvement; CGI-S = Clinical Global Impression-Severity; C-SSRS = Columbia Suicide Severity Rating Scale; DAI-10 = Drug Attitude Inventory; ECG = electrocardiogram; EOS = end of study; ET = early termination; HAMD-17 = 17-Item Hamilton Depression Rating Scale; KSS = Karolinska Sleepiness Scale; MADRS = Montgomery Asberg Depression Rating Scale; MGH ATRQ = Massachusetts General Hospital Antidepressant Treatment Questionnaire; MGH-SFI = Massachusetts General Hospital Sexual Functioning Inventory; PK = pharmacokinetic; SAS = Simpson-Angus Scale; SCID-5-CT = Structured Clinical Interview for DSM-5, Clinical Trials Version; SF-12 = 12-Item Short Form Health Survey; SIS = Sheehan Irritability Scale

[a]Safety follow-up (telephone call) will occur approximately 30 days after the last dose of study drug.

[b]Height will only be measured at the Screening visit.

[c]A single 12-lead ECG can be performed any time before blood sampling or at least 30 minutes after blood sampling during clinic visits. At Screening or Baseline, ECG can be repeated once in consultation with the Medical Monitor.

[d]To include hematology, serum chemistry, prolactin levels, and urinalysis; TSH and reflex free T4 will be done at Screening only.

[e]Clinical labs (including HbA1c at Screening only) are encouraged, but not required to be completed under fasting conditions.

[f]For female study subjects of childbearing potential, a serum pregnancy test will be completed at the Screening visit; serum and urine pregnancy tests will be completed at all other scheduled time-points.

[g]PK blood draw at the Screening visit will be for the concomitant SSRI/SNRI level only. At Baseline and Weeks 1, 3, 5, 6, 8 & 10 visits pimavanserin, AC-279 and the concomitant SSRI/SNRI antidepressant levels will be evaluated; at Baseline, the PK blood draw should be completed pre-dose. When possible, PK blood samples will also be collected from subjects experiencing a serious adverse event or an adverse event leading to discontinuation, in order to assess levels of pimavanserin and the concomitant SSRI/SNRI. For all Pk samples (scheduled and unscheduled), the date and time of the last 3 doses of study drug and the concomitant SSRI/SNRI, as well as the date and time that the blood sample was drawn, should be recorded on the source document and the eCRF.

[h]At Screening, the timeframe for the C-SSRS assessment should be lifetime and the past 6 months for Suicidal Ideation, and lifetime and the past 2 years for Suicidal Behavior. At all other visits, the timeframe for C-SSRS should be since the previous visit. The relevant C-SSRS versions should be used to capture these timeframes.

3.3 Randomization

This is a double-blind, placebo-controlled, 2-stage SPCD study. At each stage, eligible subjects will be randomized to receive:

- Pimavanserin (34 mg) + an SSRI/SNRI antidepressant or
- Placebo + an SSRI/SNRI antidepressant

The actual treatment given to an individual subject will be determined by randomization schedules. Randomization will be electronically performed using an Interactive Response Technology (IRT) based randomization system.

3.3.1 Stage 1 Double-blind Treatment Period

Subjects who meet the criteria for study eligibility will continue to receive their SSRI/SNRI antidepressant at a stable dose for the duration of the study and will be randomly assigned in a 1:3 ratio to pimavanserin or placebo, respectively.

3.3.2 Stage 2 Double-blind Treatment Period

At the end of Stage 1 (Study Week 5), subjects initially randomized to placebo and who meet the non-responder criteria will be randomly assigned in a 1:1 ratio to pimavanserin or placebo. Subjects initially randomized to pimavanserin in Stage 1 will continue to receive pimavanserin through the end of Stage 2.

3.4 Blinding

This is a double-blind study. Blinding is achieved by means of identical appearance for pimavanserin and placebo. Neither the subjects nor the study personnel at the clinical sites will know which treatment is administered to each subject.

All subjects and study personnel will be blinded to the randomization into Stage 2 of the study or continuation of the treatment. From the perspective of study subjects and investigators, the study will be conducted as a seamless 10-week double-blind treatment period.

Unplanned unblinding of treatment assignment during the study is discouraged. The investigator at a site may break the blind in the event of an immediate medical emergency, where knowledge of the subject's treatment assignment (pimavanserin or placebo) must be known in order to facilitate appropriate emergency medical treatment. The Investigator must

first attempt to contact the study Medical Monitor before unblinding a subject's treatment identity in order to obtain concurrence that unblinding a subject's treatment assignment is necessary.

For the final analysis, the treatment codes for all subjects will be released to ACADIA after all subjects have completed the study and the clinical database is locked.

3.5 Determination of Sample Size

The primary efficacy endpoint is the change from Baseline to Week 5 in the Hamilton Depression Scale (17 Items) (HAMD-17) total score. For Stage 1, Baseline is Study Week 0 and Week 5 is Study Week 5. For Stage 2, Baseline is Study Week 5 and Week 5 is Study Week 10.

A total sample size of 168 evaluable subjects is estimated to provide at least 80% power at a two-sided significance level of 0.05 assuming a mean change in the HAMD-17 total score of -7.8 (SD=8) and -10 (SD=8) in Stage 1, and -2.8 (SD=6) and -6.0 (SD=6) in Stage 2 for the placebo and pimavanserin groups, respectively.

For the above calculation, the weight for each double-blind treatment stage is equal to 0.5 and it is assumed that approximately 70% of placebo subjects will be non-responders at the end of Stage 1.

Adjusting for a potential non-evaluable rate of up to 10%, approximately188 subjects will be randomized into Stage 1 (47 to pimavanserin and 141 to placebo). It is anticipated that approximately 98 subjects will be randomized into Stage 2 (49 to pimavanserin and 49 to placebo).

4. ANALYSIS SETS

Randomized Analysis Sets

Randomized Analysis Set for Stage 1 – The Randomized Analysis Set for Stage 1 (RAS1) will consist of all subjects who were randomized into Stage 1.

Randomized Analysis Set for Stage 2 – The Randomized Analysis Set for Stage 2 (RAS2) will consist of all subjects who were placebo non-responders from Stage 1 and who were randomized into Stage 2.

Subjects will be classified according to the randomized treatment assignment.

Safety Analysis Sets

Safety Analysis Set for Stage 1 – The Safety Analysis Set for Stage 1 (SAS1) will consist of all subjects who were randomized into Stage 1 and received at least one dose of study drug in Stage 1.

Safety Analysis Set for Stage 2 – The Safety Analysis Set for Stage 2 (SAS2) will consist of all subjects who were placebo non-responders from Stage 1, were randomized into Stage 2, and received at least one dose of study drug in Stage 2.

Subjects will be classified according to the actual treatment received.

Full Analysis Sets

Full Analysis Set for Stage 1 – The Full Analysis Set for Stage 1 (FAS1) will consist of all subjects who were randomized into Stage 1, received at least one dose of study drug in Stage 1, and have both a Baseline (Study Week 0) value and at least one post-Baseline value for the HAMD-17 total score in Stage 1.

Full Analysis Set for Stage 2 – The Full Analysis Set for Stage 2 (FAS2) will consist of all subjects who were placebo non-responders from Stage 1, were randomized into Stage 2, received at least one dose of study drug in Stage 2, and have both a Baseline (Study Week 5) value and at least one post-Baseline value for the HAMD-17 total score in Stage 2.

Subjects will be classified according to the randomized treatment assignment.

Per-protocol Analysis Sets

Per-protocol Analysis Set for Stage 1 – The Per-protocol Analysis Set for Stage 1 (PAS1) will consist of a subset of subjects in FAS1 who do not have any protocol deviation which is considered to have substantial impact on the primary efficacy outcome. The precise reasons

for excluding subjects from PAS1 will be fully defined and documented *a priori* before the clinical database lock.

Per-protocol Analysis Set for Stage 2 – The Per-protocol Analysis Set for Stage 2 (PAS2) will consist of a subset of subjects in FAS2 who do not have any protocol deviation which is considered to have substantial impact on the primary efficacy outcome. The precise reasons for excluding subjects from PAS2 will be fully defined and documented *a priori* before the clinical database lock.

Subjects will be classified according to the randomized treatment assignment.

Pharmacokinetics Analysis Set for Stage 1 – The Pharmacokinetics Analysis Set 1 will consist of subjects in SAS1 with at least one measurable pimavanserin plasma concentration.

Pharmacokinetics Analysis Set for Stage 2 – The Pharmacokinetics Analysis Set 2 will consist of subjects in SAS2 with at least one measurable pimavanserin plasma concentration.

Subjects will be classified according to the actual treatment received.

5. DATA HANDLING CONVENTIONS

All data collected in the study will be listed.

5.1 General Data Reporting Conventions

Continuous variables will be summarized using the following descriptive statistics: number of subjects, mean, median, standard deviation, standard error, minimum, and maximum. Unless specified otherwise, means, medians, and confidence intervals will be presented to one more decimal place than the raw data, and the standard deviations and standard errors will be presented to two more decimal places than the raw data.

Categorical and count variables will be summarized by the number of subjects and the percent of subjects in each category. Categories with zero counts will not have zero percentages displayed. Percentages will be presented with one decimal place.

Duration in months will be calculated as ([the number of days / 365.25] *12).

Unless specified otherwise, all statistical tests will be 2-sided hypothesis tests performed at the significance level of 5% for main effects and all confidence intervals will be 2-sided 95% confidence intervals. P-values will generally be presented to 4 decimal places; values less than 0.0001 will be presented as <0.0001.

5.2 Derived Efficacy Variables

In general, assessment total scores and subscores will be derived within the analysis datasets. In the event that total scores and/or subscores are also collected on the electronic case report from (eCRF), the derived values will be used for all analyses. Both the raw and derived scores will be presented in listings.

5.2.1 Hamilton Rating Scale for Depression – 17 Items

The HAMD-17 includes a structured interview guide developed by Bech and colleagues and will be administered by the Investigator or designee based on an assessment of a subject's symptoms. This structured interview has been validated in the Danish University Antidepressant Group (DUAG) studies for use with time frames shorter than 1 week. The time frame for this scale is the past 7 days.

The HAMD-17 consists of 8 items with a score on a 3 point scale and 9 items with a score on a 5 point scale. The total score ranging from 0 to 52 will be calculated as the sum of the scores for all 17 items. A higher total score indicates more severe depression. Missing item scores will be imputed and rounded to the nearest integer as follows:

- Missing scores for items with a score on a 3 point scale will be imputed using the arithmetic mean of the non-missing scores for items with a score on a 3 point scale

- Missing scores for items with a score on a 5 point scale will be imputed using the arithmetic mean of the non-missing scores for items with a score on a 5 point scale

The total score will be considered as missing if there are missing scores for 4 or more items.

5.2.2 Sheehan Disability Scale (SDS)

The SDS is a 3-item subject-facing questionnaire used to evaluate impairments in the domains of work, social life/leisure, and family life/home responsibility. Subjects rate each item using an 11-point scale ranging from 0 (not at all) to 10 (extremely). The mean score ranging from 0 to 10 will be calculated as the arithmetic mean of the scores for all 3 items and rounded to the nearest 2 decimal places. A higher mean score indicates greater disability. If a subject indicates that no work/school occurred, the mean score will be calculated as the arithmetic mean of the other two item responses (Social Life and Family Life/Home Responsibilities) and rounded to the nearest 2 decimal places. If either of the other two item responses is missing, or if the work/school response is missing and the subject does not indicate that no work/school occurred, the mean score will be missing.

The mean scores calculated based on only two item responses (Social Life and Family Life/Home Responsibilities) will be flagged in data listings. These results will be considered observed values for data summaries and analyses.

5.2.3 Clinical Global Impressions – Severity (CGI-S) and Clinical Global Impressions – Improvement (CGI-I) Scales

The CGI-S and CGI-I are scales used by the Investigator or designee to rate the severity of the disorder and the global improvement since beginning of the study. The CGI-S rates the severity of a subject's depression over the past 7 days and the CGI-S score ranges from 1 to 7:

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

The CGI-I rates the change in a subject's depression over the past 7 days relative to the subject's symptoms at Baseline and the CGI-I score ranges from 1 to 7:

1 = Very much improved
5 = Minimally worse

2 = Much improved
6 = Much worse

3 = Minimally improved
7 = Very much worse

4 = No change

Higher CGI-S and CGI-I scores denote more severe depression and less improvement in depression respectively. Missing CGI-S and CGI-I scores will not be imputed.

5.2.4 12-Item Short Form Health Survey (SF-12V2)

The SF-12V2 assesses overall health and well-being over the past 4 weeks using Likert scale items with 3 to 5 responses.

Prior to calculating the scale scores, items #5, #6a, and #6b will be rescaled as 6 minus the original response. This effectively rescales those responses so that a response of 1 is recoded to 5 and a response of 5 is recoded to 1, etc. In addition, item #1 will be rescaled as follows: 1 recoded to 5.0, 2 recoded to 4.4, 3 recoded to 3.4, 4 recoded to 2.0, and 5 recoded to 1.0.

After the necessary rescaling, the following raw scale scores are calculated as follows:

- Physical Functioning: sum of responses to items #2a and #2b
- Role Physical: sum of responses to items #3a and #3b
- Bodily Pain: response to item #5
- General Health: response to item #1
- Vitality: response to item #6b
- Social Functioning: response to item #7
- Role Emotional: sum of responses to items #4a and #4b
- Mental Health: sum of responses to items #6a and #6c

Each raw scale is then transformed to a 0-100 scale (transformed raw score) by applying 100*(raw scale score – lowest possible raw score)/(possible raw score range).

Scale	Lowest Possible Raw Score	Possible Raw Score Range
Physical Functioning	2	4
Role Physical	2	8
Bodily Pain	1	4

General Health	1	4
Vitality	1	4
Social Functioning	1	4
Role Emotional	2	8
Mental Health	2	8

The transformed raw scores are then converted to standardized scale scores (Z-scores) using the following equations, based on the 1998 general U. S. population.

Z-score	Equation
Physical Functioning	(Transformed Raw Score – 81.18122)/29.10588
Role Physical	(Transformed Raw Score – 80.52856)/27.13526
Bodily Pain	(Transformed Raw Score – 81.74015)/24.53019
General Health	(Transformed Raw Score – 72.19795)/23.19041
Vitality	(Transformed Raw Score – 55.59090)/24.84380
Social Functioning	(Transformed Raw Score – 83.73973)/24.75775
Role Emotional	(Transformed Raw Score – 86.41051)/22.35543
Mental Health	(Transformed Raw Score – 70.18217)/20.50597

The aggregate physical and mental summary scores are then calculated using the standardized scale scores and coefficients from the 1990 general U. S. population.

Aggregate Physical Score: = (Physical Functioning * 0.42402) + (Role Physical * 0.35119) + (Bodily Pain * 0.31754) + (General Health * 0.24954) + (Vitality * 0.02877) + (Social Functioning * -0.00753) + (Role Emotional * -0.19206) + (Mental Health * -0.22069).

Aggregate Mental Score: = (Physical Functioning * -0.22999) + (Role Physical * -0.12329) + (Bodily Pain * -0.09731) + (General Health * -0.01571) + (Vitality * 0.23534) + (Social Functioning * 0.26876) + (Role Emotional * 0.43407) + (Mental Health * 0.48581).

Each aggregate score and standardized scale score is then transformed into a norm-based score (T-score) by multiplying the aggregate or standardized score by 10 and adding the resulting product to 50. The norm-based scores are distributed with a mean of 50 and a standard deviation of 10. A higher score is indicative of a better health state. Note: The T-score for Aggregate Physical Score and Aggregate Mental Score is called Physical Component Summary (PCS) score and Mental Component Summary (MCS) score respectively.

QualityMetric Health Outcomes™ Scoring Software 5.0 will be used to verify the norm-based T-scores calculated using the above formula.

5.2.5 Drug Attitude Inventory (DAI-10)

The DAI-10 contains 6 items (1, 3, 4, 7, 9, and 10) that a subject who is fully adherent to the prescribed medication would answer as "True", and 4 items (2, 5, 6, and 8) that a subject who is fully adherent to the prescribed medication would rate as "False." A correct answer is scored +1 and an incorrect answer is scored -1. The total score ranging from -10 to 10 (increment by 2) is the sum of pluses and minuses. A positive total score indicates a positive subjective response (adherent) and a negative total score indicates a negative subjective response (non-adherent). Higher scores denote better adherence. The total score will be considered as missing if at least one item score is missing.

5.2.6 Karolinska Sleepiness Scale (KSS)

The KSS is a scale that measures the subject's drowsiness and is frequently used in studies measuring subjective sleepiness. Scoring is based on a 9-point verbally anchored scale going from "1 = extremely alert" to "9 = very sleepy, great effort to keep awake, fighting sleep". Higher scores denote more drowsiness. Missing scores will not be imputed.

5.2.7 Massachusetts General Hospital Sexual Functioning Index (MGH-SFI)

Sexual functioning will be assessed using the MGH-SFI at Baseline (Study Week 0) and the MGH-SFI (follow-up version) at Study Weeks 5 and 10. MGH-SFI is a questionnaire that quantifies sexual dysfunction into 5 functional domains (“interest in sex,” “sexual arousal,” “ability to achieve orgasm,” “ability to maintain erection” [males only], and “sexual satisfaction”). Subjects rate each item using a 6-point scale ranging from 1 (greater than normal) to 6 (totally absent). MGH-SFI (follow-up version) asks one more question (in addition to the 5 functional domains in MGH-SFI) on “overall improvement since your last medication change”. Subjects rate this item using a 6-point scale ranging from 1 (very much improved) to 6 (much worse). Higher scores denote worse sexual dysfunction.

The MGH-SFI score to be analyzed will be calculated as the arithmetic mean of the item scores for the 5 domains and rounded to the nearest 2 decimal places. For MGH-SFI at Baseline (Study Week 0), the arithmetic mean will be calculated over all 5 domains for males and the 4 domains for females. For MGH-SFI (follow-up version) at Study Weeks 5 and 10, the arithmetic mean will be calculated in the same way as MGH-SFI at Baseline excluding the item score for the additional domain of “overall improvement since your last medication change”. The MGH-SFI arithmetic mean score will be considered as missing if at least one item score is missing.

5.2.8 Barratt Impulsiveness Scale (BIS-11)

The BIS-11 is a questionnaire designed to assess the personality/behavioral construct of impulsiveness. It is composed of 30 items describing common impulsive or non-impulsive (reverse scored items: 1, 7, 8, 9, 10, 12, 13, 15, 20, 29, and 30) behaviors and preferences. Items are scored on the following 4-point scale: Rarely/Never = 1; Occasionally = 2; Often = 3; Almost Always/Always = 4. For reverse scored items, a response of 1 is recoded to 4; 2 is recoded to 3; 3 is recoded to 2; and 4 is recoded to 1. The BIS-11 score ranging from 30 to 120 will be calculated as the sum of the scores for all 30 items. Higher scores denote more impulsiveness. Missing item scores will be imputed with the arithmetic mean of the non-missing item scores and rounded to the nearest integer. The BIS-11 score will be considered as missing if there are missing scores for 7 or more items.

5.2.9 Sheehan Irritability Scale (SIS)

The SIS is a 7-item subject-reported outcome measure that was developed to measure the frequency, severity, and impairment associated with irritability in psychiatric subjects. It includes items on: irritability, frustration, edginess/impatience/overreaction, moodiness, anger with self, anger with others, and temper. The recall period is the past 7 days. Items are answered on an 11-point rating scale where higher scores indicate greater severity (0= not at all, 10= extremely). The total score ranging from 0 to 70 will be calculated as the sum of the scores for all 7 items. Missing item scores will be imputed with the arithmetic mean of the non-missing item scores and rounded to the nearest integer. The total score will be considered as missing if there are missing scores for 2 or more items.

5.2.10 Montgomery Asberg Depression Rating Scale (MADRS)

The MADRS is a 10-item, clinician-rated instrument measuring depression severity. It is scored on a fixed scale of 7 points (0-6) following a structured clinician interview. The total score ranging from 0 to 60 will be calculated as the sum of the scores for all 10 items. Higher scores reflect more severe symptomatology. The total score will be considered missing if there are at least one missing item scores.

5.3 Analysis Visit Windows

Baseline for Stage 1 of the study is defined as the last non-missing result, including results from repeated and unscheduled measurements, before first dosing in Stage 1. Baseline for subjects randomized into Stage 2 is defined as the results at Study Week 5.

Efficacy, safety, and PK assessments will be summarized by analysis visit as presented in Table 2 below. For the Week 5 visit window only, the assessments collected on the date of randomization for the subjects who are randomized into Stage 2 will be selected for the analysis even if the Stage 2 randomization date is outside of the Study Day Interval, 33 - 39.

Table 2 Analysis Visit Windows

Analysis Visit	**Study Visit**	**Target Study Day @**	**Study Day Interval**
Baseline (Day 1 for Stage 1)	2 (Week 0)	1	≤ 1
Week 1	3 (Week 1)	8	2 – 11
Week 2	4 (Week 2)	15	12 – 18
Week 3	5 (Week 3)	22	19 – 25
Week 4	6 (Week 4)	29	26 – 32
Week 5	7 (Week 5)	36	33 – 39
Week 6	8 (Week 6)	43	40 – 46
Week 7	9 (Week 7)	50	47 – 53
Week 8	10 (Week 8)	57	54 – 60
Week 9	11 (Week 9)	64	61 – 67
Week 10	12 (Week 10)	71	68 – 78
Follow-up	~30-Day Follow-Up	101	≥79

@ Derivation of study day: study day = assessment date - first dose date + 1 if the assessment date ≥ first dose date, otherwise study day = assessment date – first dose date. Study day 1 is the day of first administration of study drug (pimavanserin or placebo).

Efficacy and safety assessments for Stages 1 and 2, when presented side by side or separately, will be summarized by analysis visit as presented in Table 3.

Table 3 Analysis Visit for Side-by-side or Separate Presentation of Stages 1 and 2

Data

Analysis Visit	Study Visit in Stage 1	Study Visit in Stage 2
Baseline (Day 1)	2 (Week 0)	7 (Week 5)
Week 1	3 (Week 1)	8 (Week 6)
Week 2	4 (Week 2)	9 (Week 7)
Week 3	5 (Week 3)	10 (Week 8)
Week 4	6 (Week 4)	11 (Week 9)
Week 5	7 (Week 5)	12 (Week 10)

5.3.1 Unscheduled Assessments

Both Scheduled and Unscheduled assessments, including the assessments at early termination visits, will be used for planned timepoint analyses. All assessments will be presented in data listings.

5.3.2 Multiple Measurements within Visit Windows

In the event that more than one assessment falls within a given window the assessment closest to the target study day will be selected for the by-visit analysis. If two assessments are equidistant from the target study day then the chronologically last assessment will be used. Exceptions may be made for incomplete assessments, in which case, more complete assessments may be given priority. Details are provided in a separate programming conventions document.

For safety analyses where the extreme values should be selected (e.g. overall post-Baseline minimum, overall post-Baseline maximum, and overall post-Baseline potentially clinically important values), all non-missing post-Baseline values should be considered, regardless of whether the value is selected for the by-visit summaries. All assessments will be presented in data listings.

5.4 Missing or Incomplete Date for Last Dose of Study Drug

In the Safety Analysis Set, if the last dose date of study drug is missing for a subject who completed or early terminated from the study, then the date of the end-of-study/early termination visit will be used in the calculation of treatment duration. For the incomplete last dose date of study drug, the imputation algorithms will be detailed in the analysis dataset specification document. The missing or incomplete dates will be displayed in the data listings as reported on the eCRF rather than the imputed dates.

5.5 Missing or Incomplete Dates for Prior or Concomitant Medications

Missing or incomplete medication start or stop dates will be imputed for the purpose of determining whether the medication is taken concomitantly or not (see Section 11 for definition). When the chronological order of medication use relative to the study drug treatment period is unclear due to missing or incomplete date(s), the medication will be considered as concomitant. The imputation algorithms will be detailed in the analysis dataset specification document. The missing or incomplete dates as captured on the eCRF will be displayed in the data listings.

5.6 Missing or incomplete Date for Adverse Events

Missing or incomplete adverse event (AE) start dates will be imputed for the purpose of determining whether the AEs are treatment-emergent or not (see Section 14.1 for definition). When the chronological order of an AE onset relative to the study drug treatment period is unclear due to missing or incomplete date(s), the AE will be considered as treatment-emergent. The imputation algorithms will be detailed in the analysis dataset specification document. The missing or incomplete dates captured on the eCRF will be displayed in the data listings.

5.7 Missing Severity Assessment for Adverse Events

If the severity is missing for an AE starting on or after the date of the first dose of study drug, then a severity of “Severe” will be assigned. The imputed values for severity assessment will be used for incidence summaries, while the actual values will be used in data listings.

5.8 Missing Relationship to Study Drug for Adverse Events

If the relationship to study drug is missing for an AE starting on or after the date of the first dose of study drug, a causality of “Related” will be assigned. The imputed values for

relationship to study drug will be used for incidence summaries, while the actual values will be presented in data listings.

5.9 Character Values of Clinical Laboratory Variables

If the reported value of a clinical laboratory variable cannot be used in a statistical analysis due to, for example, a character string reported for a numeric variable, an appropriately determined coded value may be used in the statistical analysis. The coding algorithms will be detailed in the analysis dataset specification document. The actual values as reported in the database will be presented in data listings.

6. SUBJECT DISPOSITION

For subjects who participate in the screening phase but are not randomized (screen failures), their demographics information (including age, sex, and primary race), screen failure reasons (the specific inclusion/exclusion criterion (or criteria) not met or other reason) and protocol version will be listed. If a subject is re-screened, then the re-screening subject ID and the final enrollment status (whether eventually enrolled) will also be displayed in this listing. In addition, the frequency that the screen failure reasons are reported will also be summarized. Note that one subject may be deemed ineligible for multiple inclusion/exclusion criteria and may be allowed to rescreen with the permission of the Medical Monitor, provided the screen failure was due to a temporary condition that subsequently resolved.

Subject disposition will be summarized to show the number of subjects screened, randomized by treatment group and stage, received study drug by treatment group and stage, discontinued by reason and treatment group and stage, and completed study by treatment group and stage.

Reasons for discontinuation will be summarized by treatment arm and overall for RAS1 and RAS2; SAS1 and SAS2; and FAS1 and FAS2. A listing will be provided displaying all subjects excluded from the Safety, Full or Per-protocol analysis sets, and will include reason(s) for exclusion.

The number of subjects screened, the number of unique subjects screened, the number of subjects randomized into Stage 1, and the number of subjects randomized into Stage 2 will be summarized overall and by site. The number of subjects enrolled at each site will also be tabulated by treatment arm and overall for RAS1 and RAS2; SAS1 and SAS2; FAS1 and FAS2; and PAS1 and PAS2.

7. PROTOCOL DEVIATIONS

Protocol deviations will be reviewed periodically over the course of the study. The review process, definition of the deviation categories, and the classification of a deviation as major or minor are detailed in the Protocol Deviation Management Plan.

Major protocol deviations will be assigned to 2 stages for the summary of major protocol deviations by stage. If a major protocol deviation happens or is identified on or after the Stage 2 randomization date or the Study Week 5 visit date for those who were not randomized into Stage 2, the major protocol deviation will be assigned to Stage 2; else it will be assigned to Stage 1.

A summary of the number and percentage of subjects with major protocol deviations by deviation category and treatment group will be provided for RAS1 (up to Study Week 5/Early termination) and RAS2 (from Study Week 5 to Study Week 10/Early termination). A listing of protocol deviations by site and subject will also be provided.

8. DEMOGRAPHICS AND OTHER BASELINE CHARACTERISTICS

Demographics and baseline characteristics will be summarized by treatment group and overall for RAS1, RAS2, SAS1, SAS2, FAS1, FAS2, PAS1, and PAS2 using descriptive statistics. Variables include age, age group (18-40 years and > 40 years), sex, race, ethnicity, height, weight, BMI, current smoker, highest education level, marital status, employment status, Montgomery Asberg Depression Rating Scale (MADRS) total score, Baseline HAMD-17 total score, Baseline HAMD-17 total score < 24 and ≥ 24, Baseline SDS mean score, and Baseline Clinical Global Impressions-Severity Illness (CGI-S).

Race will also be categorized by White vs. Non-White. The reported age reflects a subject's age at the informed consent date. Baseline CGI-S will be summarized as both a continuous variable and a categorical variable.

Depression history will be summarized by treatment group for RAS1, RAS2, SAS1, SAS2, FAS1, FAS2, PAS1, and PAS2 using descriptive statistics. Variables include:

- Age at first onset of depression symptoms
- Age at MDD diagnosis
- Duration of the current MDE (months relative to the informed consent date)
- Number of depression episodes during the subject's lifetime
- Number of hospitalizations for depression during the subject's lifetime

Depression history as described above, date of last hospitalization (if number of hospitalizations is greater than 0), and onset date for the current episode of depression will also be presented in a data listing.

Massachusetts General Hospital Antidepressant Treatment Response Questionnaire (MGH ATRQ) examines a patient's antidepressant treatment history and determines the adequacy of treatment response. MGH ATRQ is collected at Screening visit. Drug type, generic name, took during this current episode of depression (Yes/No), dose, and took at least this dose for at least 8 weeks (Yes/No), and the amount of improvement in percentage will be listed.

9. MEDICAL HISTORY

Medical history reported terms will be coded with Medical Dictionary for Regulatory Activities (MedDRA), version 19.0 or newer. The subject incidence will be summarized for each system organ class (SOC) and preferred term by treatment group and overall for FAS1, FAS2, SAS1, and SAS2. A subject will be counted only once per SOC or per preferred term for the summary.

A listing of the SOC, preferred term, body system, verbatim for the medical history condition/event, start and stop dates (when available), and an indicator for whether or not the condition is ongoing will be provided.

10. EXTENT OF EXPOSURE AND TREATMENT COMPLIANCE

Extent of exposure and treatment compliance will be summarized as continuous variables by treatment group for SAS1 in Stage 1 (5 weeks), SAS2 in Stage 2 (5 weeks), and the subjects in SAS1 who were not randomized to the pimavanserin group in Stage 2 (10 weeks).

10.1 Exposure to Study drug

Duration of exposure to study drug will be calculated for each subject as (last dose date – first dose date + 1).

For SAS1 and SAS2, the number and percentage of subjects within each of the following exposure levels (a maximum of 5 weeks [approximately 35 days]) in terms of duration (days) of exposure will also be tabulated: ≤10, 11-≤20, 21-≤30, and >30. For the PIM subjects vs. the placebo subjects in SAS1 who were not randomized to the pimavanserin group in Stage 2, the exposure levels will be: ≤20, 21-≤40, 41-≤60, and >60. Kaplan-Meier curves of duration on study drug will also be presented for each treatment group.

10.2 Measurement of Treatment Compliance

Study drug dosing compliance for a specified period is defined as the total number of tablets actually taken by a subject during that period divided by the number of tablets expected to be taken during the same period multiplied by 100. The total number of tablets actually taken is calculated by the total number of tablets dispensed minus the number of tablets returned. The number of tablets expected to be taken for a specified period is calculated as (last dose date – first dose date + 1) x 2 (the planned number of tablets taken per day).

Treatment compliance will be summarized as both a continuous variable and a categorical variable. For the categorical presentation, the number and percentage of subjects within each of the following compliance levels will be tabulated: <40%, 40 - <60%, 60 - <80%, 80 - ≤120%, and >120%.

11. PRIOR AND CONCOMITANT MEDICATION

Prior medication is defined as any medication with the start and stop dates prior to the date of the first dose of study drug. Concomitant medication is defined as any medication with a start date prior to the date of the first dose of study drug and continuing past the first dose of study drug or with a start date between the dates of the first and last doses of study drug, inclusive. Any medication with a start date after the date of the last dose of study drug will be considered as post-treatment medication. Medications will be coded using WHO Drug Dictionary March 2016 or newer version.

Concomitant medication will be assigned to 2 stages for the summary of concomitant medication by stage. A concomitant medication may be assigned to both Stage 1 and Stage 2.

The concomitant medication will be assigned to Stage 1 if the start date of any concomitant medication is prior to or on the Stage 2 randomization date (or the Study Week 5 visit date for those who were not randomized into Stage 2).

The concomitant medication will be assigned to Stage 2 if the stop date of any concomitant medication is on or after the Stage 2 randomization date (or the Study Week 5 visit date for those who were not randomized into Stage 2).

Please refer to Section 5.5 for the imputation of missing or incomplete medication start or stop dates. In general, if missing or incomplete dates make the determination impossible, the medication will be considered as concomitant and the concomitant medication will be assigned to both Stage 1 and Stage 2.

Psychiatric Medication History

The number and percentage of subjects taking psychiatric medications collected on the Psychiatric Medication History CRF pages will be tabulated by Anatomical/Therapeutic/Chemical (ATC) Level 3, preferred term, treatment group, and overall for SAS1 and SAS2 separately. Multiple medication usage by a subject in the same category will be counted only once. Listing of all psychiatric medications will be provided.

Prior and Concomitant Medications

Prior and concomitant medications will be summarized separately. The number and percentage of subjects will be tabulated by ATC Level 3, preferred term, treatment group, and overall for SAS1 (up to Study Week 5 for concomitant medications) and SAS2 (from Study Week 5 to Study Week 10 for concomitant medications) separately. Multiple medication usage by a subject in the same category will be counted only once. Listing of all prior and concomitant medications will be provided.

Concomitant medications and the incidences will also be summarized for SAS1 over 10 weeks of the study by ATC Level 3, preferred term, and treatment group. The concomitant medication incidence is determined with the number of subjects exposed to a treatment (placebo vs. pimavanserin) who took a concomitant medication while on that treatment divided by the total number of subjects exposed to that treatment.

Post-Treatment Medications

Post-treatment medications will be summarized for SAS1 by ATC Level 3, preferred term, and the study drug (placebo vs. pimavanserin) of which subjects received their last doses.

12. EFFICACY ANALYSES

All efficacy analyses will be performed using the planned treatment assignments for FAS1 and FAS2. Sensitivity analyses of the primary and key secondary efficacy endpoints will be performed using the planned treatment assignments for PAS1 and PAS2.

For Stage 1, Baseline is Study Week 0 and Week 5 is Study Week 5. For Stage 2, Baseline is Study Week 5 and Week 5 is Study Week 10.

Summary statistics for all the efficacy endpoints will be provided by visit and treatment group for FAS1 in Stage 1 (5 weeks), FAS2 in Stage 2 (5 weeks), and the subjects in FAS1 who were not randomized to the pimavanserin group in Stage 2 (10 weeks). Summary statistics for all the efficacy endpoints will also be provided by treatment group in Stage 2 for FAS2 from Baseline (Study Week 0) to Study Week 5.

12.1 Efficacy Endpoints

Primary Efficacy Endpoint

The primary efficacy endpoint is the change from Baseline to Week 5 in the HAMD-17 total score.

Key Secondary Efficacy Endpoint

The key secondary efficacy endpoint is the change from Baseline to Week 5 in the SDS mean score.

Other Secondary Efficacy Endpoints

Other secondary efficacy endpoints are the following:

- Treatment response and remission rates at the end of 5-week treatment period
- Change from Baseline to Week 5 in CGI-S score
- CGI-I score at Week 5
- Change from Baseline to Week 5 in SF-12 score
- Change from Baseline to Week 5 in DAI-10 score
- Change from Baseline to Week 5 in KSS score

- Change from Baseline to Week 5 in MGH-SFI score
- Change from Baseline to Week 5 in BIS-11 score
- Change from Baseline to Week 5 in SIS score

Treatment response is defined as a reduction from Baseline in HAMD-17 total score of 50% or more. Remission is defined as a HAMD-17 total score less than or equal to 7.

12.2 Adjustment for Covariates

The corresponding baseline value will be included as a covariate for the analysis of HAMD-17 and SDS mean scores using the mixed model for repeated measures (MMRM) as described in Sections 13.1 and 13.2.

12.3 Handling of Missing Data

The primary analyses of the primary and key secondary efficacy endpoints will be performed in 2 steps and the missing data issue is only involved in Step 1:

1. Data from Stages 1 and 2 will be analyzed separately assuming missing at random (MAR) using the stage-specific, direct likelihood-based MMRMs and scores that are missing, after any imputation of individual missing items as described in Sections 5.2.1 and 5.2.2, will not be imputed. The MMRM method is unbiased under the MAR assumption and can be thought of as aiming to estimate the treatment effect that would have been observed if all subjects had continued on treatment for the full study duration (EMA, 2009).

2. Weighted combination of statistics from the stage-specific MMRMs will be used to perform the hypothesis testing.

12.4 Multiple Comparisons / Multiplicity

The hypothesis testing for the primary and key secondary efficacy endpoints will be performed in a sequential order. That is, if there is no evidence to show the superiority of the pimavanserin treatment to the placebo with respect to the primary efficacy endpoint at the two-sided significance level of 0.05, no further testing for the key secondary efficacy endpoint will be performed. Thus, there is no need to adjust the alpha for the two comparisons between the 2 treatment groups and the family-wise Type I error rate is controlled strongly for the hypothesis testing for the primary and key secondary efficacy endpoints. Unadjusted p-values will be reported and the testing sequence will be used to determine statistical significance.

12.5 Examination of Subgroups

Treatment effect will be examined with respect to the primary and key secondary efficacy endpoints within the subgroups of sex (male and female), race (white and non-white), age (18-40 years and > 40 years), and the severity of depression at Baseline (Baseline HAMD-17 total score < 24 and ≥ 24).

13. METHODS OF EFFICACY ANALYSES

13.1 Primary Efficacy Analysis

13.1.1 Primary Analysis

The primary efficacy endpoint is the change from Baseline to Week 5 in the Hamilton Depression Scale (17 Items) (HAMD-17) total score.

Let Δ_1 and Δ_2 be the difference in the mean change from Baseline to Week 5 in the HAMD-17 total score between the pimavanserin and placebo groups in Stage 1 and Stage 2 respectively. Let *w* be the weight for Stage 1 and 1-*w* the weight for Stage 2.

The null hypothesis is: $w\Delta_1 + (1-w)\Delta_2 = 0.$

The alternative hypothesis is: $w\Delta_1 + (1-w)\Delta_2 \neq 0.$

The hypothesis testing will be performed using the weighted combination of statistics from the stage-specific MMRM. The MMRMs will include effects for treatment group, visit, treatment-by-visit interaction, Baseline HAMD-17 total score, and the Baseline HAMD-17 total score-by-visit interaction. An unstructured covariance matrix will be used and the Kenward-Roger approximation will be used to adjust the denominator degrees of freedom.

In the event that the model fails to converge using the unstructured covariance matrix, the following covariance structures will be modeled in the order given (i.e. from least parsimonious to most parsimonious): heterogeneous Toeplitz, heterogeneous compound symmetry, heterogeneous autoregressive(1), Toeplitz, compound symmetry, autoregressive(1), variance components. The first covariance structure that allows for convergence will be selected for the final model.

The overall treatment effect will be assessed as the weighted differences between the pimavanserin and placebo groups in least-squares mean change from Baseline to Week 5. The prespecified weight *w*, is 0.5/0.5 for Stage 1/Stage 2.

Inference will be conducted using the following weighted linear combination of stage-wise treatment effects (Chen et al., 2011 and Fava et al., 2003):

$$Z = \frac{w\hat{\Delta}_1+(1-w)\hat{\Delta}_2}{\sqrt{w^2Var(\hat{\Delta}_1)+(1-w)^2Var(\hat{\Delta}_2)}} \sim N(0,1) \quad (1)$$

In the above formula, $w = 0.5$ and $\hat{\Delta}_1$ and $\hat{\Delta}_2$ are the differences in least squares (LS) means between pimavanserin and placebo at Week 5, for Stages 1 and 2, respectively.

Summary statistics for the HAMD-17 total score (observed and change from Baseline), LS means, the between-group difference in LS mean with the corresponding 95% confidence interval, p-value, and the effect size (Cohen's d) will be presented at each post-Baseline visit for FAS1 and FAS2.

The treatment effect size (Cohen's d) is calculated using the following formula:

Effect Size = LS mean difference / SD (2)

SD is the model-based estimate, i.e., the estimated standard deviation from the unstructured covariance matrix. The sign (+ or -) of the effect size will be chosen so that a positive value favors pimavanserin.

LS mean ± SE from the MMRM models over time for the change from Baseline values by treatment group will be displayed in line plots for FAS1 in Stage 1 and FAS2 in Stage 2.

The weighted treatment differences in LS means between pimavanserin and placebo, the corresponding 95% CI, and the p-value will be presented at Week 5.

13.1.2 Sensitivity Analyses

The following 2 sensitivity analyses of HAMD-17 total score are planned.

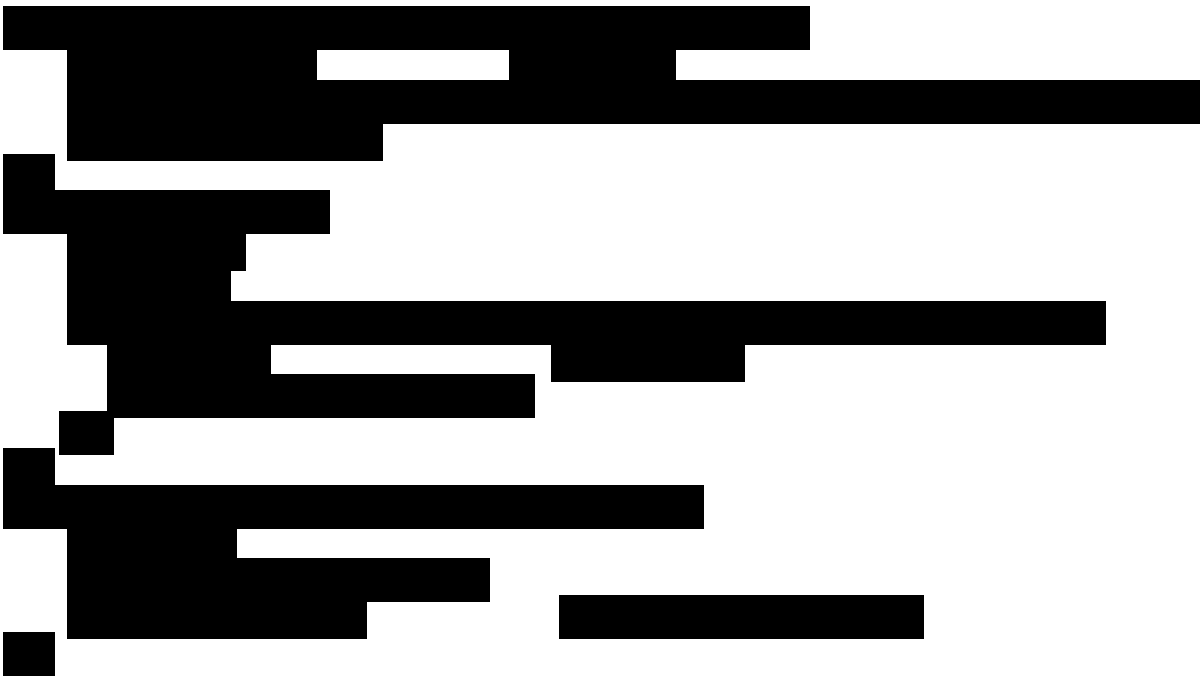
1. Per-protocol Analysis

A sensitivity analysis similar to the primary analysis will be performed for PAS1 and PAS2.

2. Pattern-Mixture Models Assuming Missing Not At Random (MNAR)

The sensitivity analysis is implemented using multiple imputations that are based on the distribution of placebo group responses over time. The underlying assumption is that subjects with missing data follow the distribution of the placebo responses. The following 3 steps are involved for both FAS1 and FAS2 separately:

- The posterior mean and covariance estimates from the SAS MI procedure using the available non-missing placebo data and random number seeds of 1030421 for Stage 1 and 103042 for Stage 2 will be utilized to impute missing data in the HAMD-17 total score. The imputed values will be constrained to be within the range of 0 to 52. [redacted]



- The change from Baseline to Week 5 in the HAMD-17 total score will then be calculated and analyzed for each of the 100 fully imputed datasets using an ANCOVA model with treatment group as a factor and the Baseline value as a covariate.

- The treatment LSMEAN differences will be averaged and the associated standard errors will be summarized based on within-imputation and between-imputation variance using the SAS MIANALYZE procedure, to yield a combined estimate for treatment effect with its associated 95% CI and p-value.

Inference will be made using the weighted linear combination of stage-wise, combined estimates for treatment effect using Equation (1).

13.2 Key Secondary Efficacy Analysis

The key secondary efficacy endpoint is the change from Baseline to Week 5 in the SDS mean score. The hypotheses, the primary analysis method, and the line plots for the key secondary efficacy endpoint are the same as the hypotheses, the primary analysis method, and line plots for the primary efficacy endpoint described in Section 13.1.1 but with SDS replacing HAMD-17. A per-protocol analysis similar to the primary analysis will be performed for PAS1 and PAS2.

13.3 Other Secondary Efficacy Analyses

ANCOVA analyses for the secondary efficacy endpoints in this section will be performed using (1) the observed values (i.e., missing data will not be imputed) and (2) the last observation carried forward (LOCF) method for missing data (Note: for subjects without post-Baseline values, the Baseline values will be carried forward).

Remission rate at the end of the two 5-week treatment periods

Remission is defined as a HAMD-17 total score less than or equal to 7. Missing HAMD-17 total scores will be considered as non-remissions and the subjects with missing HAMD-17 total scores will be included in the denominator of the remission rate.

The remission rate will be compared between the pimavanserin and placebo groups within each stage using Pearson's chi-square test.

Let n be the total number of subjects in FAS1, n_{1q} be the number of placebo subjects in FAS1, n_{1p} be the number of pimavanserin subjects in FAS1, n_{2q} be the number of placebo subjects in FAS2 (i.e., placebo non-responders in Stage 1, were randomized to the placebo group in Stage 2, received at least one dose of study drug in Stage 2, and have both a Baseline [Study Week 5] value and at least one post-Baseline value for the HAMD-17 total score in Stage 2), and n_{2p} be the number of pimavanserin subjects in FAS2 (i.e., placebo non-responders in Stage 1, were randomized to the pimavanserin group in Stage 2, received at least one dose of study drug in Stage 2, and have both a Baseline [Study Week 5] value and at least one post-Baseline value for the HAMD-17 total score in Stage 2).

Let r_{1q} be the number of placebo remission subjects in FAS1 in Stage 1, r_{1p} be the number of pimavanserin remission subjects in FAS1 in Stage 1, r_{2q} be the number of placebo remission subjects in FAS2 in Stage 2, and r_{2p} be the number of pimavanserin remission subjects in FAS2 in Stage 2.

Let $q_1 = r_{1q} / n_{1q}$ and $p_1 = r_{1p} / n_{1p}$ be placebo and pimavanserin remission rate in Stage 1. Let $q_2 = r_{2q} / n_{2q}$ and $p_2 = r_{2p} / n_{2p}$ be placebo and pimavanserin remission rate in Stage 2, among

placebo non-responders in Stage 1. Let *w* be the weight for Stage 1 and 1-*w* be the weight for Stage 2 (Note: *w* is 0.5 for this study).

The null hypothesis is: $w\,(P_1 - Q_1) + (1 - w)\,(P_2 - Q_2) = 0.$

The alternative hypothesis is: $w\,(P_1 - Q_1) + (1 - w)\,(P_2 - Q_2) \neq 0.$

Inference will be made using the following weighted linear combination of stage-wise treatment effects:

$$Z = \frac{w(p_1 - q_1) + (1 - w)(p_2 - q_2)}{\sqrt{w^2 Var(p_1 - q_1) + (1 - w)^2 Var(p_2 - q_2)}} \sim N(0, 1),$$

where, $Var(p_i - q_i) = \frac{p_i(1-p_i)}{n_{ip}} + \frac{q_i(1-q_i)}{n_{iq}},\ i = 1, 2.$ (3)

Treatment response rate at the end of the two 5-week treatment periods

Treatment response is defined as a reduction from Baseline in HAMD-17 total score of 50% or more. The hypotheses and the analysis method for treatment response rate are the same as the hypotheses and the analysis method for the remission rate but with treatment response replacing remission.

Change from Baseline to Week 5 in CGI-S score

The change from Baseline to Week 5 in CGI-S score will be analyzed in a similar way as the primary analysis of the primary efficacy endpoint using the weighted linear combination test with the Baseline HAMD-17 total score replaced by the Baseline CGI-S score.

CGI-I score at Week 5

The CGI-I score will be analyzed in a similar way as the primary analysis of the primary efficacy endpoint using the weighted linear combination test. The differences are (1) the response is CGI-I score (as opposed to the change from Baseline in CGI-I score) and (2) the Baseline CGI-S score will be used as the covariate in the MMRM models.

The CGI-I data will also be dichotomized by combining “very much improved” and “much improved” into 1 category (“Improved”), and the remaining items into the other (“Not Improved”). For analysis purposes, missing CGI-I scores and CGI-I scores of 0 (“not assessed”) will be treated as “Not Improved”.

The observed score and dichotomized data will be descriptively summarized by treatment group and visit. Treatment comparisons with respect to the dichotomized data within each stage for FAS1 and FAS2 will be performed using Pearson's chi-square test. No overall treatment comparisons combining Stages 1 and 2 will be made with respect to the dichotomized data.

Change from Baseline to Week 5 in SF-12 scores

The change from Baseline to Week 5 in the norm-based Aggregate Physical Score and the norm-based Aggregate Mental Score will be summarized by treatment group and analyzed within each stage for FAS1 and FAS2 using the analysis of covariance (ANCOVA) model with treatment group as a factor and the corresponding Baseline value as a covariate. The overall treatment effect combining the 2 stages will be tested in a similar way as the primary efficacy endpoint using the weighted linear combination test in Equation (1).

Change from Baseline to Week 5 in DAI-10 score

The change from Baseline to Week 5 in DAI-10 score will be analyzed in a similar way as the SF-12 scores are analyzed.

Change from Baseline to Week 5 in KSS score

The change from Baseline to Week 5 in the 9-point KSS scale score will be analyzed in a similar way as the primary analysis of the primary efficacy endpoint using the weighted linear combination test with the Baseline HAMD-17 total score replaced by the Baseline KSS score.

Change from Baseline to Week 5 in MGH-SFI score

The change from Baseline to Week 5 in MGH-SFI score will be analyzed in a similar way as the SF-12 scores are analyzed.

Change from Baseline to Week 5 in BIS-11 score

The change from Baseline to Week 5 in BIS-11 score will be analyzed in a similar way as the primary analysis of the primary efficacy endpoint using the weighted linear combination test with the Baseline HAMD-17 total score replaced by the Baseline BIS-11 score.

Change from Baseline to Week 5 in SIS score

The change from Baseline to Week 5 in SIS score will be analyzed in a similar way as the primary analysis of the primary efficacy endpoint using the weighted linear combination test with the Baseline HAMD-17 total score replaced by the Baseline SIS score.

13.4 Exploratory Efficacy Analyses

No planned exploratory efficacy analyses are defined.

14. SAFETY **ANALYSES**

All safety analyses including the analysis of plasma concentration data will be performed using the actual treatment for SAS1 in Stage 1 (5 weeks) and SAS2 in Stage 2 (5 weeks). To evaluate treatment differences in 10-week safety data, unless otherwise specified, safety data from Baseline (Study Week 0) to Study Week 10 will also be summarized by treatment group for subjects in SAS1 who were not randomized to the pimavanserin group in Stage 2.

For Stage 1, Baseline is Study Week 0 and Week 5 is Study Week 5. For Stage 2, Baseline is Study Week 5 and Week 5 is Study Week 10.

14.1 Adverse Events

Adverse events will be coded using MedDRA dictionary, Version 19.0 or newer.

An AE (classified by preferred term) will be considered a treatment-emergent AE (TEAE) if started after first study dose administration and no later than last study dose date + 30. AEs reported on Day 1 based on Baseline (pre-dose) findings (e.g. clinically significantly abnormal vital signs, laboratory test results, or electrocardiogram parameters) will not be considered as TEAEs.

TEAEs will be classified into 2 stages for the summary of TEAEs by stage. If the onset date of a TEAE is on or after the Stage 2 randomization date or the Study Week 5 visit date for those who were not randomized into Stage 2, the TEAE will be assigned to Stage 2; else it will be assigned to Stage 1.

The number and percentage of subjects reporting TEAEs in each treatment group will be tabulated by system organ class (SOC) and preferred term; by SOC, preferred term, and maximum severity. The number and percentage of subjects reporting related TEAEs in each treatment group will also be tabulated by SOC and preferred term. If more than 1 AE occurs with the same preferred term for the same subject, then the subject will be counted only once for that preferred term using the most severe and most related occurrence for the summarization by severity and by relationship to study drug.

TEAEs will also be tabulated by treatment group and without displaying the SOC terms; this table will be sorted in descending order of frequency of preferred term within the pimavanserin group.

The incidence of most frequently reported (preferred terms reported by ≥ 5% of subjects in any treatment group) TEAEs for Stages 1 and 2 only, treatment emergent SAEs, and TEAEs leading to discontinuation of study drug will be summarized by SOC, preferred term, and treatment group. The tables will be sorted alphabetically by SOC and then by descending

order of subjects frequency within the pimavanserin group. In addition, the incidence of fatal treatment-emergent AEs (i.e., events that cause death) will be summarized separately by preferred term and treatment group.

An AE listing by subject will display all events, including those which occur during screening, and will include the verbatim term in addition to the MedDRA SOC and preferred term. This listing will also include all relevant eCRF data associated with the event: date of onset, date resolved, date of last dose, severity, frequency, outcome, relationship to study drug, and action taken with study drug. Separate listings will be presented for subjects with treatment-emergent SAEs, subjects with TEAEs leading to discontinuation and subject who died (if any).

14.2 Clinical Laboratory Variables

Clinical laboratory assessments are performed at Screening Visit 1, Baseline (Study Week 0), Study Week 5, and Study Week 10/Early termination.

- Clinical chemistry serum tests include the following:
    - Sodium (Na), potassium (K), chloride (Cl), phosphorus (P), calcium (Ca), carbon dioxide (CO_2), blood urea nitrogen (BUN), creatinine (CR), uric acid
    - Alanine aminotransferase (ALT/SGPT), aspartate aminotransferase (AST/SGOT), gamma-glutamyl transpeptidase (GGT), alkaline phosphatase (ALP), total bilirubin (TBIL), lactate dehydrogenase (LDH)
    - HbA1c (Screening only)
    - Glucose
    - Albumin (ALB), total protein
    - Prolactin
    - Creatine kinase (CK)/creatine phosphokinase (CPK)
    - Lipid panel
        - Total cholesterol, HDL-cholesterol, triglycerides, LDL-cholesterol, Cholesterol/HDL ratio, Non-HDL cholesterol
- Hematology tests include the following:

- Complete blood count (CBC) including:
  - White blood cell (WBC) count
  - Complete differential (relative and absolute)
  - Hematocrit (Hct), hemoglobin, red blood cells (RBC), platelets
  - Reticulocyte count

- Urinalysis tests include the following:
  - Blood, RBCs, WBCs, protein, glucose, ketones, specific gravity, pH

Clinical laboratory values (in Système International [SI] units) and the change from Baseline values will be summarized by treatment group at each post-Baseline visit using descriptive statistics. The overall minimum and maximum post-Baseline observed and change from Baseline values will also be summarized. For hemoglobin, hematocrit and uric acid, the above summaries will be presented for each gender as well as for both genders combined. For urinalysis with categorical results, the number and percentage of subjects will be tabulated by category at Baseline, Study Week 5 and Study Week 10, and the denominator is the number of subjects with non-missing values for the given parameter, visit and treatment group.

Laboratory values will also be summarized in shift tables by treatment group, to determine the number and percentage of subjects with values classified as below, within, and above normal ranges at each post-Baseline visit relative to the same classification at the Baseline visit. For the by-visit shit summary, the denominator is the number of subjects with non-missing values at Baseline and the given visit for the given parameter and treatment group. For the shift to the overall post-Baseline minimum or maximum, all post-Baseline values will be considered, including unscheduled and out of window values and the denominator is the number of subjects with non-missing Baseline value and at least 1 post-Baseline value for the given parameter and treatment group. For hemoglobin, hematocrit and uric acid, the shift summaries will be presented for each gender as well as for both genders combined.

Clinical laboratory values are potentially clinically important (PCI) if they meet either the low or high PCI criteria listed in Tables 4 and 5. The number and percentage of subjects with post-Baseline PCI values for each of the categories in Table 4 and 5 will be summarized by treatment group for selected parameters. For the overall post-Baseline summaries of PCI values, all post-Baseline values will be considered, including unscheduled and out of window values. Subjects with multiple PCI values for a given parameter will be counted only once for that parameter. For the overall post-Baseline summary, the numerator of the percentage is the number of subjects with at least 1 post-Baseline PCI laboratory value for the given parameter and treatment group, and the denominator is the number of subjects with at least 1 post-

Baseline laboratory value for the given parameter and treatment group. For hemoglobin, hematocrit and uric acid, the count and percentage of subjects with PCI values will be presented for each gender as well as for both genders combined. Subjects with PCI values will be presented in an additional listing.

Table 4 Criteria for Potentially Clinically Important Laboratory Values – Hematology and Chemistry

Analyte	Conventional Unit	Low PCI Criteria	High PCI Criteria	SI Unit	Low PCI Criteria	High PCI Criteria
Hematology (whole blood)						
Hemoglobin (male)	g/dL	<11	>18	g/L	<110	>180
Hemoglobin (female)	g/dL	<10	>17	g/L	<100	>170
Hematocrit (male)	%	<30	>55	L/L	<0.3	>0.55
Hematocrit (female)	%	<30	>50	L/L	<0.3	>0.5
Leukocyte (White Blood Cell Count)	x 10^3/uL	≤2.8	≥15	x 10^9/L	≤2.8	≥15
Neutrophils	x 10^3/uL	≤1.5	No upper limit	x 10^9/L	≤1.5	No upper limit
Platelet Count	x 10^3/uL	≤75	≥700	10^9/L	≤75	≥700
Chemistry (serum or plasma)						
ALT (SGPT)	U/L	No lower limit	≥3 X ULN	U/L	No lower limit	≥3 X ULN
AST (SGOT)	U/L	No lower limit	≥3 X ULN	U/L	No lower limit	≥3 X ULN
Total Bilirubin	mg/dL	No lower limit	≥1.5 ULN	umol/L	No lower limit	≥1.5 ULN
BUN	mg/dL	No lower limit	≥30.0	mmol/L	No lower limit	≥10.71
Creatine Kinase (CK)	U/L	No lower limit	≥3 ULN	U/L	No lower limit	≥3 ULN
Sodium	mEq/L	≤125	≥155	mmol/L	≤125	≥155
Potassium	mEq/L	≤3.0	≥5.5	mmol/L	≤3.0	≥5.5
Calcium, total	mg/dL	<8.0	>11.0	mmol/L	<2.0	>2.75
Lactate Dehydrogenase (LDH)	U/L	No lower limit	≥3 X ULN	U/L	No lower limit	≥3 X ULN
Alkaline Phosphatase	U/L	No lower limit	≥3 X ULN	U/L	No lower limit	≥3 X ULN
Uric acid (male)	mg/dL	No lower limit	≥10.5	umol/L	No lower limit	≥624.75
Uric acid (female)	mg/dL	No lower limit	≥8.5	umol/L	No lower limit	≥505.75
Albumin	g/dL	≤2.6	≥6.0	g/L	≤26	≥60

Analyte	Conventional Unit	Low PCI Criteria	High PCI Criteria	SI Unit	Low PCI Criteria	High PCI Criteria
Total Protein	g/dL	≤5.0	≥10.0	g/L	≤50	≥100
Chloride	mEq/L	≤85	≥120	mmol/L	≤85	≥120
Glucose (random)	mg/dL	≤45.1	≥200.0	mmol/L	≤2.48	≥11
Serum Creatinine	mg/dL	Not Applicable	>1.5 ULN	umol/L	Not Applicable	>1.5 ULN
Triglycerides	mg/dL	Not Applicable	>300	mmol/L	Not Applicable	>3.39
Gamma-Glutamyl Transferase (GGT)	U/L	Not Applicable	≥3 ULN	U/L	Not Applicable	≥3 ULN

Table 5 Criteria for Potentially Clinically Important Laboratory Values - Urinalysis

Urinalysis (qualitative dipstick)	Low PCI Criteria	High PCI Criteria
Blood (Occult Blood)	Not Applicable	≥+2
Protein	Not Applicable	≥+2
Glucose	Not Applicable	≥+2

Clinical laboratory data will be displayed in data listings with date and study day of collection. All units will be displayed according to SI conventions for units. Out of range values will be flagged in the data listings (i.e., 'L' or 'H'). A separate listing will be provided for a subset of the chemistry, hematology, and urinalysis analytes with values classified as PCI.

The pregnancy results (positive or negative) for female subjects will be presented in a listing.

14.3 Vital Signs

Vital signs including weight and the derived BMI will be collected throughout the study at every visit. Observed vital signs and the changes from Baseline at each post-Baseline visit will be summarized by treatment group using descriptive statistics.

Vital sign values will be considered PCI if they meet the criteria listed in Table 6. The number and percentage of subjects with post-Baseline vital signs that are PCI will be summarized by treatment group at each post-Baseline visit and for overall post-Baseline. For the overall post-Baseline summaries, all post-Baseline values will be considered, including unscheduled and out of window values. Subjects with multiple PCI values for a given parameter will be counted only once for that parameter. For the by-visit summary, the numerator for the percentage is the number of subjects with a post-Baseline PCI vital sign for the given parameter, visit and treatment group, and the denominator is the number of subjects with non-missing values for the given parameter, visit and treatment group. For the overall post-

Baseline summary, the numerator for the percentage is the number of subjects with at least 1 post-Baseline PCI vital sign for the given parameter and treatment group, and the denominator is the number of subjects with at least 1 post-Baseline vital sign for the given parameter and treatment group. A listing of subjects with any PCI value will be provided.

Table 6 Criteria for Potentially Clinically Important (PCI) Vital Signs

Vital Sign Parameter	Unit	Criteria			
		Observed Value	**And/Or**	**Change Relative to Baseline**	**Change from Supine to Standing**
Systolic blood pressure (supine or sitting)	mmHg	≥180	And	Increase of ≥20	-
		≤90	And	Decrease of ≥20	-
Diastolic blood pressure (supine or sitting)	mmHg	≥105	And	Increase of ≥15	-
		≤50	And	Decrease of ≥15	-
Pulse (supine or sitting)	bpm	≥120	And	Increase of ≥15	-
		≤50	And	Decrease of ≥15	-
Weight	kg	Not Applicable		Increase of ≥7%	-
				Decrease of ≥7%	-

14.4 Electrocardiogram (ECG)

12-lead ECGs are collected at Screening Visit 1, Baseline (Study Week 0), Study Week 1, Study Week 5, Study Week 6, and Study Week 10/Early termination. Observed values of ECG variables (e.g., heart rate, PR interval, QRS interval, QT interval, and QTc interval) and the changes from Baseline at each assessment time point will be summarized by treatment group.

QTcF will also be categorized into the following categories (msec) and the number and percentage of subjects in each category will be summarized by treatment group at each visit and for the overall post-Baseline maximum:

- Observed: ≤450, 451 - ≤480, 481 - ≤500, and >500; >450; >480.
- Change from Baseline: ≤10, 11 – 30, 31 – 60, and >60; >30.

Electrocardiogram variable values will be considered PCI if they meet the criteria listed in Table 7. The number and percentage of subjects with post-baseline PCI values will be summarized by treatment group at each post-Baseline visit and for overall post-Baseline. For the by-visit summary, the numerator for the percentage is the number of subjects with a post-Baseline PCI ECG for the given parameter, visit and treatment group, and the denominator is the number of subjects with non-missing values for the given parameter, visit and treatment group. For the overall post-Baseline summary, the numerator for the percentage is the number of subjects with at least 1 post-Baseline PCI ECG for the given parameter and treatment group, and the denominator is the number of subjects with at least 1 post-Baseline ECG value for the given parameter and treatment group. A listing of all subjects with any PCI value will be provided.

Table 7 Criteria for Potentially Clinically Important ECG Values

ECG Parameter	Unit	High PCI Criteria
QRS Interval	msec	≥120
PR Interval	msec	≥220
QTcB or QTcF	msec	>500
QTcB or QTcF: change from baseline	>60 msec	

14.5 Physical Examination

Physical examinations are performed at Screening Visit 1, Baseline (Study Week 0), Study Week 5, and Study Week 10. Physical examination results (normal, abnormal, and not done) will be summarized in a frequency table by treatment group, body system and visit.

14.6 Other Safety Endpoints

Unless otherwise specified, scores derived based on sub-scores for the safety endpoints in this section will be considered missing if any corresponding sub-scores are missing. No imputations will be performed. Summary statistics for all safety questionnaire data discussed in this section will also be provided by treatment group in Stage 2 for SAS2 from Baseline (Study Week 0) to Study Week 5.

Columbia-Suicide Severity Rating Scale (C-SSRS)

The C-SSRS Baseline/Screening version will be completed at the Screening visit and the version assessing information since the last visit will be completed at all following visits (including the Baseline visit).

There are 5 questions about suicidal ideation, representing 5 types of suicidal ideation: wish to be dead; non-specific active suicidal thoughts; active suicidal ideation with any methods (not plan) without intent to act; active suicidal ideation with some intent to act, without specific plan; active suicidal ideation with specific plan and intent. If a subject answers "yes" to any of these 5 questions at any post-Baseline visit including unscheduled and out of window visits, this subject will be counted as having suicidal ideation in this study.

There are 5 questions about suicidal behavior, representing 5 types of suicidal behavior: actual attempt; interrupted attempt; aborted attempt; preparatory acts or behavior; suicide. If a subject answers "yes" to any of these 5 questions at any post-Baseline visit including unscheduled and out of window visits, this subject will be counted as having suicidal behavior in this study.

Suicidality is defined as a subject who reported at least 1 occurrence of suicidal ideation or at least 1 occurrence of suicidal behavior at any post-Baseline visit including unscheduled and out of window visits.

All data will be listed. The event counts and the number and percentage of subjects reporting any post-Baseline suicidal ideation, suicidal behavior, or suicidality will be summarized by treatment group for SAS1 over Stage 1 (5 weeks), SAS2 over Stage 2 (5 weeks), and the subjects in SAS1 who were not randomized to the pimavanserin group in Stage 2 over the whole study (10 weeks).

Abnormal Involuntary Movement Scale (AIMS)

The AIMS is a rating scale that was designed in the 1970s to measure involuntary movements known as tardive dyskinesia. The AIMS has a total of 12 items rating involuntary movements of various areas of the patient's body. These items are rated on a 5-point scale of severity. The scale is rated from 0 (none), 1 (minimal), 2 (mild), 3 (moderate), 4 (severe). Two of the 12 items refer to dental care. The remaining 10 items refer to body movements themselves. Total AIMS scores range from 0 to 42.

Observed total AIMS score and the changes from Baseline at each post-Baseline visit will be summarized by treatment group using descriptive statistics.

The number and percentage of subjects with dyskinesia will be summarized by treatment group at each visit and for overall post-Baseline for SAS1 over Stage 1 (5 weeks), SAS2 over Stage 2 (5 weeks), the subjects in SAS1 who were not randomized to the pimavanserin group in Stage 2 over the whole study (10 weeks), and SAS2 over Stage 1 (5 weeks). Dyskinesia is defined as having a score of 3 or more on any of the first 7 AIMS items or a score of 2 or more on any two of the first 7 AIMS items. The tabulations will be presented for subjects who have at least 1 AIMS assessment as well as for a subset of these subjects who do not have dyskinesia at Baseline.

The individual item scores will be listed but not summarized.

Barnes Akathisia Scale (BARS)

The BARS is a clinician-rated scale to assess drug-induced akathisia and classify it as absent, mild, moderate, or severe. The BARS consists of items that assess the objective presence and frequency of akathisia, the level of an individual's subjective awareness and distress, and global severity. Objective Akathisia, Subjective Awareness of Restlessness, and Subjective Distress Related to Restlessness are rated on a 4-point scale from 0 – 3 and are summed yielding a total score ranging from 0 to 9. The Global Clinical Assessment of Akathisia uses a 6-point scale ranging from 0 – 5.

Observed total score, the score of Global Clinical Assessment of Akathisia, and the corresponding changes from Baseline at each post-Baseline visit will be summarized by treatment group using descriptive statistics.

The number and percentage of subjects with akathisia will be summarized by treatment group at each visit and for overall post-Baseline for SAS1 over Stage 1 (5 weeks), SAS2 over Stage 2 (5 weeks), the subjects in SAS1 who were not randomized to the pimavanserin group in Stage 2 over the whole study (10 weeks), and SAS2 over Stage 1 (5 weeks). Akathisia is defined as having a Global Clinical Assessment of Akathisia score ≥2. The tabulations will be presented for subjects who have at least 1 BARS assessment as well as for a subset of these subjects who do not have akathisia at Baseline.

The individual item scores will be listed but not summarized. .

Simpson Angus Scale (SAS)

The SAS is composed of 10 items and is used to assess pseudoparkinsonism. The grade of severity of each item is rated using a 5-point scale. Total SAS scores can range from 0-40. Signs assessed include gait, arm-dropping, shoulder shaking, elbow rigidity, wrist rigidity, leg pendulousness, head dropping, glabella tap, tremor, and salivation. .

Observed total SAS score and the changes from Baseline at each post-Baseline visit will be summarized by treatment group using descriptive statistics.

The number and percentage of subjects with Parkinsonism will be summarized by treatment group at each visit and for overall post-Baseline for SAS1 over Stage 1 (5 weeks), SAS2 over Stage 2 (5 weeks), the subjects in SAS1 who were not randomized to the pimavanserin group in Stage 2 over the whole study (10 weeks), and SAS2 over Stage 1 (5 weeks). Parkinsonism is defined as having a SAS total score >3. The tabulations will be presented for subjects who have at least 1 SAS assessment as well as for a subset of these subjects who do not have Parkinsonism at Baseline.

The individual item scores will be listed but not summarized.

For the overall post-Baseline summaries, all post-Baseline values will be considered, including unscheduled and out of window values.

15. CLINICAL PHARMACOKINETIC AND PHARMACODYNAMIC ANALYSES

For pimavanserin-treated subjects, plasma concentration for pimavanserin and AC-279 will be listed. Plasma concentration data for pimavanserin and AC-279 will be summarized at each visit using descriptive statistics. Concentrations that are below the limit of quantification (BLQ) will be displayed as "BLQ" in the data listings and imputed as 0 for computing summary statistics.

Plasma concentration data for the concomitant SSRIs/SNRIs will be listed.

If data allow, population PK and PK/PD analyses will be performed to further characterize the PK profile and exposure response relationship of pimavanserin and its metabolite using measures of safety, and efficacy parameters. The results of population PK and PK/PD modeling will be presented in a separate report. Pimavanserin plasma concentration data will remain blinded until the unblinding of the clinical database at the end of the study.

16. INTERIM ANALYSIS

No interim analysis is planned in this study.

17. DATA MONITORING/REVIEW COMMITTEE

There is no Data Monitoring Committee in this study.

18. COMPUTER METHODS

Statistical analyses will be performed using Version 9.4 (or newer) of SAS® (SAS® Institute, Inc., Cary, North Carolina) on a suitably qualified and validated environment.

Validation and quality control of the tables, listings and figures containing the results of the statistical analyses will follow appropriate standard operating procedures.

19. CHANGES TO ANALYSES SPECIFIED IN PROTOCOL

No changes are made to the analyses specified in the protocol.

20. REFERENCES

Chen Y.F., Yang Y., Hung H., Wang S. (2011). Evaluation of performance of some enrichment designs dealing with high placebo response in psychiatric clinical trials. *Contemporary Clinical Trials,* 32, 592-604.

EMA (2009). *Guideline on Missing Data in Confirmatory Clinical Trials, European Medicines Agency*, London, UK.

Fava M., Evins A.E., Dorer D.J., Schoenfeld D.A. (2003). The problem of the placebo response in clinical trials for psychiatric disorders: culprits, possible remedies, and a novel study design approach. *Psychotherapy and Psychosomatics,* 72, 115–127.

21. APPENDICES

21.1 Summary of Version Changes

Version No:	Document History Description of Update	Author(s)	Version Date
Final	Original version	[redacted]	21AUGUST2018