Statistical Analysis Plan for Study M16-837

Phase 2, Multicenter, Double-blind (Sponsor-unblinded), Randomized, Placebo-Controlled Study of the Safety and Efficacy of Elagolix in Women with Polycystic Ovary Syndrome

Date: 28 October 2019

Version 1.0

Table of Contents

List of Tables

List of Figures

List of Appendices

1.0 Introduction

This Statistical Analysis Plan (SAP) describes the statistical analyses for Elagolix Study M16-837 'Phase 2, Multicenter, Double-blind (Sponsor-unblinded), Randomized, Placebo-Controlled Study of the Safety and Efficacy of Elagolix in Women with Polycystic Ovary Syndrome.'

The analyses of pharmacokinetic endpoints, pharmacodynamic endpoints, and biomarkers will not be covered in this SAP.

The SAP will not be updated in case of administrative changes or amendments to the protocol unless the changes impact the analysis.

Unless noted otherwise, all analyses will be performed using SAS Version 9.3 (SAS Institute Inc., Cary, NC 27513) or later under the UNIX operating system.

2.0 Study Design and Objectives

2.1 Objectives

The objective of the study is to assess the pharmacokinetics, pharmacodynamics, safety, and efficacy of elagolix in women with Polycystic Ovary Syndrome (PCOS).

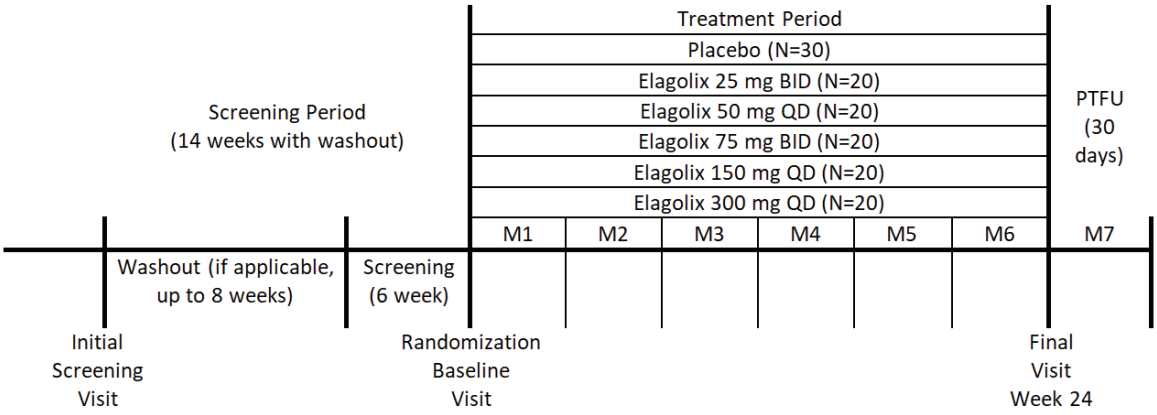
2.2 Study Design Overview

This is a Phase 2, randomized, double-blind (sponsor-unblinded), 6-month, placebo controlled, parallel group, multicenter study designed to assess the PK and PD of elagolix and to demonstrate the safety and efficacy of elagolix in women with PCOS.

The duration of the study could be up to 42 weeks, which includes an 8-week Washout (if necessary), a 6-week Screening Period, a 24-week Treatment Period, and a 30-day Post-Treatment Follow-up Period.

The schematic of the study is presented in Figure 1.

Figure 1. Study Schematic



BID = twice daily; M = month; PTFU = post-treatment follow-up; QD = once daily

2.3 Treatment Assignment and Blinding

Stratified randomization will be performed by a ratio of 2:2:2:2:2:3 to the following treatments: 25 mg BID, 50 mg QD, 75 mg BID, 150 mg QD, 300 mg QD, or placebo (20 subjects in each active treatment group and 30 subjects in the placebo group). Subject randomization will be stratified based on Ferriman-Gallwey (FG) score ($< 8, \geq 8$) and body mass index (BMI) ($< 30, \geq 30$).

2.4 Sample Size Determination

The planned total sample size is 130 subjects: 20 subjects per elagolix treatment group and 30 subjects assigned to placebo.

The sample size will provide at least 85% power to detect a difference between each elagolix group and the placebo group in the proportion of menstrual cycle responders based on a Pearson Chi-square test, assuming response rates of 10% for the placebo group and 50% for the elagolix group with 2-sided $\alpha = 0.05$. The above sample size was calculated using nQuery advisor 7.0.

3.0 Endpoints

3.1 Primary Endpoint(s)

The primary endpoint is the proportion of menstrual cycle responders as defined in Section 8.3.1.

3.2 Secondary Endpoint(s)

The secondary endpoint is change from Baseline to Week 1 in the area under the luteinizing hormone (LH) serum concentration-time curve (AUC) as defined in Section 8.4.

3.3 Other Efficacy Endpoint(s)

The primary and secondary efficacy endpoints are listed in Section 3.1 and Section 3.2, respectively. The additional efficacy endpoints are:

- Proportion of complete menstrual cycle responders (normal menstrual cycles by the beginning of Month 3 and maintained thereafter each month through Month 6).
- Change from Baseline in the LH serum AUC at each scheduled assessment during the Treatment Period.
- Change from Baseline in the FSH serum AUC at each scheduled assessment during the Treatment Period.
- Change from Baseline in total and free testosterone serum concentration at each scheduled assessment during the Treatment Period.
- Number of menstrual bleeding episodes during the Treatment Period.
- Number of ovulations based on weekly serum progesterone level during the Treatment Period.
- Proportion of subjects with at least 2 ovulations during the Treatment Period
- Endpoints related to hyperandrogenism – Patient Global Impression (PGI) for both acne and hirsutism; change from Baseline at each scheduled assessment during the Treatment Period for the following: acne measured with total lesion

count and investigator's global assessment (IGA), hirsutism measured with Ferriman–Gallwey (FG) score.

- Endpoints related to metabolic features - change from Baseline at each scheduled assessment during the Treatment Period for the following: glycated hemoglobin (HbA1c), fasting insulin/glucose, body mass index (BMI), lipids, waist circumference.
- Change from Baseline at each scheduled assessment during the Treatment Period for additional hormones: androstenedione, anti-Mullerian hormone (AMH), estradiol (E2), progesterone.
- Change from Baseline in ovarian volume at each scheduled assessment during the Treatment Period.
- Endpoints related to the subject's quality of life: change from Baseline in domains from the Polycystic Ovary Syndrome Questionnaire (PCOSQ) (emotions, body hair, weight, infertility problems, and menstrual problems) and the Menstrual Bleeding Questionnaire (MBQ) at each scheduled assessment during the Treatment Period.

3.4 Safety Endpoint(s)

Safety evaluations include physical examination, vital signs, pelvic ultrasound (transabdominal ultrasound [TAU] and transvaginal ultrasound [TVU]), clinical laboratory tests (including hematology, chemistry, urinalysis, and lipid panel), bone mineral density (BMD) via dual energy X-ray absorptiometry (DXA) examinations, and adverse event (AE) monitoring for the entire study duration.

4.0 Analysis Populations

The following population sets will be used for the analyses.

The Full Analysis Set (FAS) consists of all randomly assigned subjects who have received at least one dose of study drug in this study. The data from the FAS will be presented by the treatment group assigned at the time of randomization, even if the subject does not

receive the correct treatment or does not follow the protocol until completion. The FAS will be used for all baseline and efficacy analyses.

The Safety Analysis Set consists of all subjects who received at least one dose of study drug. The data from the safety analysis set will be presented by the treatment actually received no matter what treatment group was assigned at the time of randomization. If a subject takes more than one treatment, the subject will be analyzed in the safety analysis set as taking the treatment to which she was randomized. All safety analyses will be performed based on the safety analysis set.

5.0 Subject Disposition

The total number of subjects who were randomized and who completed the study will be summarized by treatment group and overall. In addition, reasons for study discontinuation will be summarized.

A summary of subject accountability will be provided where the number of subjects in each of the following categories will be summarized for each treatment group:

- Subjects randomized in the study;
- Subjects who took at least one dose of study drug;
- Subjects who completed protocol-specified treatment;
- Subjects who prematurely discontinued study drug;

The number and percentage of subjects who discontinued study drug will be summarized by treatment group and overall as follows:

- By any reason for discontinuation;
- By primary reason for discontinuation.

In addition, the number and percentage of subjects who discontinued study will be summarized by treatment group and overall as follows:

- By any reason for discontinuation;
- By primary reason for discontinuation.

6.0 Study Drug Duration

For the Safety Analysis Set, duration of treatment will be summarized for each treatment group and for all investigational study drug dose groups combined. Duration of treatment is defined for each subject as last dose date minus first dose date + 1. Duration of treatment will be summarized using the number of subjects treated, mean, standard deviation, median, minimum and maximum. In addition, the number and percentage of subjects in each treatment duration interval (1 – 28, 29 – 56, 57 – 84, 85 – 112, 113 – 140, 141 – 168, and > 168 days) will be summarized.

7.0 Demographics, Baseline Characteristics, Medical History, and Prior/Concomitant Medications

Demographics, baseline or disease characteristics, medical history, and prior and concomitant medications will be summarized by treatment group and overall using the FAS. Categorical variables will be summarized with the number and percentage of subjects; percentages will be calculated based on the number of non-missing observations. Continuous variables will be summarized with descriptive statistics (number of non-missing observations, mean and standard deviation, median, minimum and maximum).

7.1 Demographics and Baseline Characteristics

Continuous demographic variables include age, weight, height, waist circumference, and body mass index (BMI). Categorical demographic variables include sex, ethnicity, race, age (18 – 26, 27 – 35 years), BMI (18.5 – < 25, 25 – < 30, ≥ 30 kg/m^2), tobacco use (current, former, never, unknown), and alcohol user (current, former, never, unknown).

Disease characteristics include menstrual and obstetrical history. Continuous variables include number of days bleeding. Categorical variables include cycle length (> 35 days

[yes/no]), typical intensity of menstrual periods (light, moderate, heavy), ever been pregnant (yes/no), and number of pregnancies (0, 1, 2, 3, 4, > 4), and congenital abnormality (yes/no).

7.2 Medical History

Medical history and gynecological medical/surgical history data will be coded using the Medical Dictionary for Regulatory Activities (MedDRA). The actual version of the MedDRA coding dictionary will be noted in the statistical tables and clinical study report. The number and percentage of subjects in each medical history category (by MedDRA system organ class and preferred term) will be summarized by treatment group and overall. The system organ class (SOC) will be presented in alphabetical order, and the preferred terms will be presented in alphabetical order within each SOC. Subjects reporting more than one condition/diagnosis will be counted only once in each row (SOC or preferred term).

7.3 Prior and Concomitant Medications

Prior and concomitant medications will be summarized by generic name.

Prior medications are those medications with a start date prior to the first study drug administration date.

Concomitant medications are those medications taken during the Treatment Period with an end date after the first dose of study drug or ongoing at the end of the study, and a start date prior to the last dose of study drug. A medication will be considered a concomitant medication where one of the following three cases occur (1) the start date is missing and the end date is either after or on the first study drug dose date; (2) the start date is not missing and the end date is missing; (3) both the start date and the end date are missing.

Prior and concomitant medications will be summarized by ATC Classification and preferred terms from the World Health Organization (WHO) Drug Dictionary with number and percentage for each treatment group and overall. A subject who reports two

or more uses of the same medication will be counted only once within each WHO preferred term. A subject with medications with more than one generic name will be counted only once in the overall total.

8.0 Efficacy Analyses

8.1 General Considerations

All efficacy analyses will be conducted using the FAS. Unless otherwise specified, each elagolix dose group will be compared to placebo. All statistical tests will be conducted at an alpha level of 0.1 (two-sided). A test will be deemed significant if the P value rounded to three decimal places is less than or equal to 0.1 unless otherwise specified.

Unless otherwise specified, categorical data will be summarized by frequencies and percentages for each treatment group and overall; continuous data will be summarized by the mean, standard deviation, median, minimum, and maximum for each treatment group and overall. For the analyses of change from baseline, the within-group changes will be summarized with the mean, standard deviation, and the median; between-group differences will be summarized with the mean, standard error, P value, and the 90% CIs.

Baseline refers to the last non-missing value obtained prior to or on Study Day 1.

8.2 Handling of Missing Data

Missing data will be imputed using the following methods for the efficacy analyses:

Mixed-Effect Model Repeat Measurement (MMRM): The repeated measures analyses will be conducted using mixed models including observed measurements at all visits. The mixed model includes the categorical fixed effects of treatment, visit and treatment-by-visit interaction, subject as a random effect, and the continuous fixed covariates of baseline measurement. An unstructured variance covariance matrix will be used. Parameter estimation is based on the assumption of data being missing at random and using the method of restrictive maximum likelihood (REML). Missing data will not be imputed for MMRM analysis.

Observed Cases (OC): The OC analysis will not impute values for missing evaluations, and thus a subject who does not have an evaluation on a scheduled visit will be excluded from the OC analysis for that visit. OC will exclude values after a subject prematurely discontinues from study drug.

Multiple Imputation (MI): The MI approach will be used as a sensitivity analysis for the primary endpoint. For the primary endpoint, 20 'complete' datasets will be generated using SAS PROC MI. The seed will be 123. Using the Cochran-Mantel-Haenszel (CMH) model adjusted for FG score ($< 8, \geq 8$) and normal/obese status (BMI $< 30, \geq 30$) at Baseline, the imputed endpoints will be analyzed using each of the 20 datasets. SAS PROC MIANALYZE will be used to generate the final inferences of difference of proportion of menstrual cycle responders between each elagolix group and placebo.

8.3 Primary Efficacy Endpoint(s) and Analyses

8.3.1 Primary Efficacy Endpoint(s)

The primary efficacy endpoint is the proportion of menstrual cycle responders.

A subject is considered as a menstrual cycle responder if she has at least 2 normal menstrual cycles during the final 4 months of the Treatment Period. For subjects who complete the treatment period this will be Month 3 to Month 6; for subjects who prematurely discontinue at Month X, where X is greater than or equal 4, this will be Month (X – 3) to Month X; for subjects who prematurely discontinue at Month X, where X is less than 4, this will be Month 1 to Month X. A normal menstrual cycle will be defined as follows:

a. Subject has menstrual bleeding (excluding spotting only) AND
b. The length of the menstrual cycle is from 21 to 35 days. Menstrual cycle length is defined as the length of time from the first day of one menstrual bleeding episode to the day before the next menstrual bleeding episode (calculated by [first day of the next menstrual bleeding episode] minus [the first day of the referenced bleeding episode]).

A menstrual bleeding episode is defined with the following rules:

a. A bleeding episode requires at least one of the following:
   - At least 2 consecutive days of bleeding, OR
   - At least 1 day of bleeding followed immediately by 1 day of spotting or 1 day of missing data then spotting or 1 day of missing data then bleeding, OR
   - At least 1 day of spotting followed immediately by 1 day of bleeding or 1 day of missing data then bleeding.

b. Consecutive days of spotting only with no days of bleeding are not considered as an episode.

c. If consecutive days of spotting are adjacent to a day with bleeding, then these consecutive days with spotting are counted as part of the same bleeding episode.

d. One day of missing data between a spotting day and bleeding day or between 2 bleeding days qualifies as part of the same bleeding episode.

Examples are presented in Table 1.

Table 1. Examples of Menstrual Bleeding Episodes

Example	Days									Menstrual Bleeding Episode (Yes/No)	Duration (days)
	1	2	3	4	5	6	7	8	9		
A	×	●	×	×	×	×	×	×	×	No	N/A
B	×	○	○	×	×	×	×	×	×	No	N/A
C	×	●	●	×	×	×	×	×	×	Yes	2
D	×	○	●	×	×	×	×	×	×	Yes	2
E	×	●		●	×	×	×	×	×	Yes	3
F	×	●	×	●	×	×	×	×	×	No	N/A
G	×	○	●	●	×	×	×	×	×	Yes	3
H	×	○		●	×	×	×	×	×	Yes	3
I	×	○			●	×	×	×	×	No	N/A
J	×	○	●	●		●	○	×	×	Yes	6
K	×	○		○	○	●	●	○	×	Yes	5

× = no spotting or bleeding; ○ = spotting; ● = bleeding; blank = missing; gray shading indicates menstrual bleeding episodes; N/A = not applicable

8.3.2 Handling of Missing Data for the Primary Efficacy Endpoint(s)

The primary analysis will be based on observed cases only; missing data will not be imputed for the primary endpoint.

8.3.3 Primary Efficacy Analysis

The 5 primary null hypotheses for this study is that there are no differences in the proportion of women that are menstrual cycle responders between each elagolix dose group and placebo. Each null hypothesis will be tested against the alternative hypothesis that there is a difference between an elagolix group and placebo.

Each elagolix group will need to demonstrate a statistically significantly greater proportion of women that are menstrual cycle responders in order for that elagolix group to be considered more efficacious than placebo for the primary endpoint. That is, an elagolix group will be considered statistically more efficacious than placebo if the

two-sided p-value for the comparison of the proportion of menstrual cycle responders between that elagolix group and placebo is less than or equal to 0.1 and in favor of the elagolix group.

For the primary efficacy endpoint, the pairwise comparisons for the difference in proportion of menstrual cycle responders between each elagolix group and placebo will be analyzed using the Cochran-Mantel-Haenszel test (CMH) adjusted for FG score ($< 8, \geq 8$) and normal/obese status (BMI $< 30, \geq 30$) at Baseline. Additionally, the CMH-based 90% confidence interval for the difference in proportion of responders within each FG score ($< 8, \geq 8$) and normal/obese status (BMI $< 30, \geq 30$) at Baseline will be calculated based on normal approximation.

8.3.4 Additional Analyses of the Primary Efficacy Endpoint(s)

The following sensitivity analyses for the primary efficacy endpoint will be conducted:

1. The menstrual cycle response for subjects who prematurely discontinue the study drug at or before Month 4 will be considered as missing. Multiple imputation (MI) approach will be used for the primary endpoint. Details regarding MI are in Section 8.2.
2. Subjects who prematurely discontinue study drug during the Treatment Period due to "lack of efficacy" or adverse event (AE) will be considered as non-responders, while subjects who discontinue study drug during the Treatment Period due to other reasons will be categorized as responders/non-responders in the same manner as done in the primary analysis.
3. All subjects who prematurely discontinue the study drug at or before Month 4 will be considered as non-responders and the primary analysis will be repeated. Subjects who discontinue after Month 4 will be categorized as responders/non-responders in the same manner as done in the primary analysis.
4. The difference between each elagolix group and placebo group in proportion of menstrual cycle responders will be analyzed using contracts within a logistic

regression model with the responder/non-responder categorization as the dependent variable, treatment as the main effect, and baseline FG score (< 8, ≥ 8) and baseline normal/obese status (BMI < 30, ≥ 30) as the covariates. The proportion of menstrual cycle responders, difference in proportions between each elagolix group and placebo, and odds ratio with corresponding 90% confidence interval and P value will be presented.

8.4 Secondary Efficacy Analyses

The key secondary endpoint is change from baseline in the area under the luteinizing hormone (LH) serum concentration-time curve over the 4 hours (area under the curve [AUC]) at Day 7. Serial blood samples for LH assay will be collected at Baseline and Weeks 1, 4, and 16 according to the following schedule: 0 hour, 0.5, 1, 1.5, 5, 2.5, 3, 3.5, and 4 hours. The AUC over the 4 hours will be calculated based on the linear trapezoidal rule.

The analysis for the key secondary endpoint will be based on a mixed-effects model with repeated measures (MMRM). If appropriate, the natural logarithmic transformation or other transformation may be employed for LH AUC. The MMRM analysis will include the fixed categorical effects of treatment, visit, treatment-by-visit interaction, baseline FG score (< 8, ≥ 8), and baseline normal/obese status (BMI < 30, ≥ 30), subject as a random effect, and the continuous fixed covariate of baseline LH AUC. Each elagolix dose group will be compared to the placebo group; the LS mean of treatment difference and associated 90% CI and p-value will be reported.

8.5 Additional Efficacy Analyses

8.5.1 Proportion of Complete Menstrual Cycle Responders

A subject is considered as a complete menstrual cycle responder if she has normal menstrual cycles by the beginning of Month 3 and maintained thereafter each month through Month 6 during the Treatment Period. Subjects who discontinued before

Month 6 during the Treatment Period will be considered as non-responders. A normal menstrual cycle has been defined in Section 8.3.1.

The pairwise comparisons for the difference in proportion of complete menstrual cycle responders between each elagolix group and placebo will be analyzed using the Cochran-Mantel-Haenszel test (CMH) adjusted for FG score ($< 8, \geq 8$) and normal/obese status (BMI $< 30, \geq 30$) at Baseline. Additionally, the CMH-based 90% confidence interval for the difference in proportion of complete menstrual cycle responders within each FG score ($< 8, \geq 8$) and normal/obese status (BMI $< 30, \geq 30$) at Baseline will be calculated based on normal approximation.

Same analysis will be conducted excluding subjects who premature discontinue during the Treatment Period.

8.5.2 LH, FSH, testosterone and other hormones

The changes from baseline to each scheduled assessment of the hormone parameters will be analyzed using a mixed-effects model with repeated measures (MMRM). If appropriate, the natural logarithmic transformation or other transformation may be employed for the hormone parameters. The MMRM analysis will include the fixed categorical effects of treatment, visit, treatment-by-visit interaction, baseline FG score ($< 8, \geq 8$), and baseline normal/obese status (BMI $< 30, \geq 30$), subject as a random effect. and the continuous fixed covariate of baseline hormone value.

8.5.3 Number of Menstrual Bleeding Episodes

Number of menstrual bleeding episodes during the Treatment Period will be analyzed using one-way ANCOVA with treatment group as the main effect, and baseline FG score ($< 8, \geq 8$) and normal/obese status (BMI $< 30, \geq 30$) as the covariates. Subjects who premature discontinue during the Treatment Period will be excluded from the analysis.

In addition, number of menstrual bleeding episodes per months of treatment will be analyzed using one-way ANCOVA with treatment group as the main effect, and baseline FG score (< 8, ≥ 8) and normal/obese status (BMI < 30, ≥ 30) as the covariates.

8.5.4 Ovulations

Determination of ovulation will be based on weekly serum progesterone level; a progesterone level ≥ 3 ng/mL will be used to indicate ovulation. One ovulation is defined as one weekly or more than one consecutive weekly serum progesterone level ≥ 3 ng/mL with the prior weekly serum progesterone level < 3 ng/mL or missing and the following weekly serum progesterone level < 3 ng/mL or missing.

Number of ovulations during the Treatment Period will be analyzed using one-way ANCOVA with treatment group as the main effect, and baseline FG score (< 8, ≥ 8) and normal/obese status (BMI < 30, ≥ 30) as the covariates. In addition, the proportion of subjects with at least 2 ovulations during the Treatment Period will be analyzed using the Cochran-Mantel-Haenszel test (CMH) adjusted for FG score (< 8, ≥ 8) and normal/obese status (BMI < 30, ≥ 30) at Baseline. Subjects who premature discontinue during the Treatment Period will be excluded from both analyses.

In addition, number of ovulations per months of treatment will be analyzed using one-way ANCOVA with treatment group as the main effect, and baseline FG score (< 8, ≥ 8) and normal/obese status (BMI < 30, ≥ 30) as the covariates.

8.5.5 Change from baseline for other clinical efficacy endpoints

Change from baseline to each scheduled assessment for the following endpoints will be analyzed: acne measured with total lesion count and investigator global assessment of acne, hirsutism measured with Ferriman–Gallwey (FG) score, metabolic features (Hb A1c, fasting insulin/glucose, body mass index (BMI), lipids [LDL cholesterol, HDL cholesterol, total cholesterol, triglycerides], waist circumference), ovarian volume, and endpoints related to subject's quality of life (PCOSQ [emotions, body hair, weight, infertility problems, and menstrual problems], and MBQ). The analysis will be based on

one-way ANCOVA with treatment group as the main effect, and baseline FG score (< 8, ≥ 8), baseline normal/obese status (BMI < 30, ≥ 30), and baseline value of the corresponding endpoint being analyzed as the covariates.

8.5.6 Patient Global Impression (PGI) for Acne and Hirsutism

For patient global impression for acne and hirsutism (PGIA and PGIH), the number and percentage of subjects in each response category will be summarized at each scheduled assessment during the Treatment Period. No statistical test will be performed.

For PGIA and PGIH, the response categories of "Much better" and "A little better" will be combined together. The remaining four categories will be combined and labeled "Otherwise." The number and percentage of subjects with response of (a) "Much better" or "A little better," and (b) "Otherwise" will be summarized at each scheduled assessment during the Treatment Period. Comparisons between each elagolix group and placebo will be made using a logistic regression model with (a) "Much better" or "A little better"/(b) "Otherwise" as the dependent variable, treatment as the main effect, and baseline FG score (< 8, ≥ 8) and baseline normal/obese status (BMI < 30, ≥ 30) as the covariates.

9.0 Safety Analyses

9.1 General Considerations

Safety analyses will be carried out using Safety Analysis Set. All safety analyses will be based on observed data. Unless otherwise specified, missing data will not be imputed.

For continuous variables, when the analyses of change and/or percent change from Baseline to post-baseline visit(s) are performed, the within-group change from Baseline to each relevant visit will be summarized by treatment group with the mean, SE, and 95% CIs. The between-group differences will be summarized with the mean, SE, 95% CIs, and P value when applicable. P value will be compared to an alpha level of 0.05. At each post-baseline visit, the Baseline mean and post-baseline visit mean will be calculated for all subjects with baseline and post-baseline value at that visit by treatment group.

Categorical data will be summarized by number and percentage of subjects by treatment group.

9.2 Adverse Events

Adverse events (AEs) will be summarized and presented using primary MedDRA System Organ Classes (SOCs) and preferred terms (PTs) according to the version of the MedDRA coding dictionary used for the study at the time of database lock. The actual version of the MedDRA coding dictionary used will be noted in the AE tables and in the clinical study report. Specific adverse events will be counted once for each subject for calculating percentages, unless stated otherwise. In addition, if the same adverse event occurs multiple times within a subject, the highest severity and level of relationship to investigational product will be reported.

9.2.1 Treatment-Emergent Adverse Events

Treatment-emergent AEs are defined as any AE with the onset that is after the first dose of study drug. Events where the onset date is the same as the study drug start date are assumed to be treatment-emergent. All treatment-emergent AEs will be summarized overall, as well as by primary MedDRA SOC and Preferred Term. The SOCs will be presented in alphabetical order, and the PTs will be presented in alphabetical order within each SOC.

The number and percentage of subjects experiencing treatment-emergent AEs will be summarized.

9.2.2 Adverse Event Overview

An overview of AEs will be presented consisting of the number and percentage of subjects experiencing at least one event for each of the following AE categories:

- Any treatment-emergent AE
- Any treatment-emergent AE related to study drug according to the investigator
- Any severe treatment-emergent AE

- Any serious treatment-emergent AE
- Any treatment-emergent AE leading to discontinuation of study drug
- Any treatment-emergent AE of special interest as specified in Appendix A
- Any treatment-emergent AE leading to death
- All deaths

9.2.3 Treatment-Emergent Adverse Events by SOC and PT

Treatment-emergent adverse events will be summarized by SOC and PT; by maximum relationship to study drug as assessed by the investigator (e.g., reasonable possibility or no reasonable possibility) and SOC and PT; by maximum severity and SOC and PT; and by subject number and SOC and PT. Specific adverse events will be counted once for each subject for calculating percentages, unless stated otherwise. In addition, if the same adverse event occurs multiple times within a subject, the highest severity and level of relationship to investigational product will be reported.

In addition, treatment-emergent adverse events will be summarized by PT and sorted by decreasing frequency for the total active group.

9.2.4 Deaths, Other Serious Adverse Events, and Adverse Events Leading to Study Drug Discontinuation

The number and percentage of subjects experiencing the following categories will be tabulated according to the primary MedDRA SOC and PT. In addition, listing of subjects experiencing the following categories will be generated.

- Treatment-emergent SAEs
- Treatment-emergent AEs leading to discontinuation of study drug
- Treatment-emergent AEs leading to death

9.2.5 Adverse Events of Special Interest

Adverse events of special interest will be summarized by SOC and PT and will be based on standardized or company MedDRA queries (SMQs or CMQs), or based on adjudication results. Adverse events of special interest are categorized as follows:

- Elevated hepatic transaminases
- BMD decrease/fractures
- Mood and depression-related events

Detailed information about the search criteria are provided in Appendix A.

9.3 Analysis of Laboratory Data

Data collected from central and local laboratories, including additional laboratory testing due to an SAE, will be used in all analyses, except for Baseline where SAE-related laboratory assessments on or before the first dose of study drug will be excluded. The clinical laboratory tests defined in the protocol operations manual (e.g., hematology and clinical chemistry) will be summarized.

Each laboratory variable will be summarized for all time points (starting with Baseline) with the number of non-missing observations, mean and standard deviation, median, minimum and maximum. Mean change from baseline to each applicable post-baseline visit will be summarized for selected laboratory variables, with the number of observations, baseline mean, and visit mean. The change from baseline mean, standard error, and 95% confidence interval will be presented for the mean change from baseline within each treatment group and difference between each elagolix group and placebo. Selected laboratory variables are lipid variables, liver variables (alkaline phosphatase, ALT, AST, bilirubin), hemoglobin, hematocrit, platelet count, and red blood cell (RBC) count.

Changes in laboratory parameters will be tabulated using shift tables either by NCI CTC criteria or categorized as low, normal, or high based on the normal ranges of the

laboratory used for each sample. A shift table from baseline either to the worse value (based on NCI CTC criteria) during treatment or to minimum and maximum value (based on normal range), will be created. A similar shift table will be provided to summarize shifts from baseline to the final post-baseline value.

Laboratory abnormalities meeting CTCAE criteria grade 3 and 4 will be summarized.

The number and percentage of subjects meeting the following criteria will be summarized at Baseline and each relevant visit by treatment group in the Treatment Period:

- Total cholesterol: ≤ 300, $> 300 – \leq 400$, $> 400 – \leq 500$, and > 500 mg/dL
- HDL-C: < 40 and ≥ 40 mg/dL
- LDL-C: < 130, $\geq 130 – < 160$, $\geq 160 – < 190$, and ≥ 190 mg/dL
- TG: ≤ 150, $> 150 – \leq 300$, $> 300 – \leq 500$, $> 500 – \leq 1000$, and > 1000 mg/dL

In addition, the number and percentage of subjects who have potentially clinically significant (PCS) lipid values meeting the following criteria any time during the Treatment Period will be summarized by treatment group: Listings will be provided to summarize subject-level laboratory data for subjects meeting PCS criteria.

- Total cholesterol > 200 mg/dL
- LDL-C > 130 mg/dL
- LDL-C > 160 mg/dL
- HDL-C < 40 mg/dL
- TG > 150 mg/dL

The number and percentage of subjects in each treatment group with maximum on treatment laboratory values meeting the following criteria compared to the upper limit of normal (ULN) will be summarized by treatment group to assess potential hepatotoxicity.

- ALT $\geq 3 \times$ ULN and total bilirubin $\geq 2 \times$ ULN
- AST $\geq 3 \times$ ULN and total bilirubin $\geq 2 \times$ ULN

- ALT and AST ≥ 3 × ULN and total bilirubin ≥ 2 × ULN
- ALT ≥ 3 × ULN and total bilirubin < 2 × ULN
- AST ≥ 3 × ULN and total bilirubin < 2 × ULN
- ALT and AST ≥ 3 × ULN and total bilirubin < 2 × ULN
- ALT ≥ 3 × ULN and total bilirubin ≥ 1.5 × ULN
- AST ≥ 3 × ULN and total bilirubin ≥ 1.5 × ULN
- ALT and AST ≥ 3 × ULN and total bilirubin ≥ 1.5 × ULN
- ALT ≥ 3 × ULN, ≥ 5 × ULN, ≥ 10 × ULN, ≥ 20 × ULN
- AST ≥ 3 × ULN, ≥ 5 × ULN, ≥ 10 × ULN, ≥ 20 × ULN
- Total bilirubin ≥ 1.5 × ULN, ≥ 2.0 × ULN

The maximum ratio relative to the ULN is used to determine if subjects met the criteria listed above. The ALT, AST, and total bilirubin values do not need to be concurrent in order to meet the defined criteria. For ALT, AST, and total bilirubin, a subject is counted if the post-baseline laboratory value during the Treatment Period meets the above criteria regardless of Baseline laboratory value (i.e., the post-baseline laboratory value does not need to be worse than Baseline laboratory value).

A listing of all ALT, AST, total bilirubin, and alkaline phosphatase values will be provided for each subject who met any of the criteria defined above.

The following plots will be generated: peak AST vs. peak bilirubin; peak ALT vs. peak bilirubin.

9.4 Analysis of Vital Signs and Weight

Vital sign variables include sitting systolic blood pressure and sitting diastolic blood pressure.

Each vital sign variable and weight will be summarized for all time points (starting with Baseline) with the number of non-missing observations, mean and standard deviation, median, minimum and maximum. Mean change from baseline to each applicable post-

baseline visit will be summarized for each vital sign variable, with the number of observations, baseline mean, and visit mean. The change from baseline mean, standard error, and 95% confidence interval will be presented for the mean change from baseline within each treatment group and difference between each elagolix group and placebo.

The number and percentage of subjects who have potentially clinically significant (PCS) vital sign and weight values meeting the following criteria will be summarized by treatment group. All increase/decrease is calculated from Baseline to a post-baseline visit in the Treatment Period.

- Diastolic blood pressure
  - ≤ 50 mmHg and ≥ 15 mmHg decrease
  - > 90 mmHg and ≥ 15 mmHg increase
  - ≥ 100 mmHg
- Systolic blood pressure
  - ≤ 90 mmHg and ≥ 20 mmHg decrease
  - ≥ 140 mmHg and ≥ 20 mmHg increase
  - ≥ 160 mmHg
- Weight
  - ≥ 5% decrease
  - ≥ 7% increase.

Listings will be provided to summarize subject-level vital sign and weight data for subjects meeting PCS criteria.

9.5 Other Safety Analyses

9.5.1 Columbia Suicide Severity Rating Scale (C-SSRS)

The C-SSRS data will be summarized as observed by treatment group.

The C-SSRS – Baseline/Screening measured at Day 1 will be considered as Baseline C-SSRS. Baseline C-SSRS will be summarized by treatment group. Other analysis of

C-SSRS will only include subjects who have at least 1 post-baseline C-SSRS measurement, regardless of whether they had a baseline C-SSRS measurement. The number of subjects with suicidal ideation, suicidal behavior, and self-injurious behavior without suicidal intent at each scheduled assessment during Treatment Period will be summarized. In addition, this table will be repeated for providing a summary of lifetime outcomes and past year outcomes at screening and Day 1. No statistical test will be performed.

The number of subjects with suicide-related treatment-emergent events based on the C-SSRS during Treatment Period will be summarized. No statistical tests will be performed.

A listing of subjects with suicidal ideation, suicidal behavior, or self-injurious behavior without suicidal intent based on the C-SSRS during treatment period will be provided.

9.5.2 Bone Mineral Density

All analyses and summaries of bone mineral density (BMD) will be performed for each segment, i.e., femoral neck, lumbar spine, and total hip. For subjects who had a right-side scan performed (rare instances), their data for right femoral neck and right femoral total hip will be included in the analysis with the data for the left femoral neck and left total hip, respectively (available for the majority of subjects). If more than one scan is reported for an anatomical segment within an analysis window, the worse (the lower value) of the multiple measurements will be used for analysis for each anatomical segment. The analyses below will exclude subjects who switch machine manufacturer types between baseline and post-baseline.

A continuous summary of BMD at Baseline and Month 6 will be provided by treatment group. This summary will include the mean, SD, median, minimum, and maximum.

The analysis of percent change from baseline in BMD during the Treatment Period will be based on an ANCOVA model with treatment as the main effect and baseline value of

corresponding parameter as a covariate. Each elagolix group will be compared to placebo.

The number and percentage of subjects with percent change from Baseline to Month 6 during the Treatment Period in mutually exclusive categories of BMD decrease: ≤ 0%, > 0% – ≤ 1.5%, > 1.5% – ≤ 3%, > 3% – ≤ 5%, > 5% – < 8%, and ≥ 8% will be summarized by treatment group. The comparison between each elagolix group and placebo will be made using Fisher's exact test.

A listing of subjects with BMD decrease from baseline ≥ 8% will be provided.

9.5.3 Pregnancy Results

Pregnancy information will be summarized by treatment group.

10.0 Overall Type-I Error Control

This is a Phase 2 proof of concept study; therefore, there will be no control of Type I error for primary, secondary, or exploratory analyses for this study.

11.0 Version History

Table 2. SAP Version History Summary

Version	Date	Summary
1.0	29 Oct 2019	Original version

12.0 References

Appendix A. Definition of Adverse Events of Special Interest

Adverse Events of Special Interest (AESI) will be identified using the following search criteria:

Item of Safety Interest	Method of Surveillance
Bone Mineral Density Loss	Osteoporosis/Osteopenia SMQ
Bone Fractures	Osteoporosis/Osteopenia SMQ
Mood and Depression-related Events	Depression and Suicide/Self-Injury SMQ
Elevated Hepatic Transaminases	Hepatic Questionnaire and the Drug-related hepatic disorders – comprehensive SMQ