STATISTICAL ANALYSIS PLAN

Version 1.0, dated 30-May-2018

A Randomized, Placebo-Controlled, Double-Blind, Dose-Ranging, Phase 2b Study to Investigate the Efficacy of ESN364 in Postmenopausal Women Suffering From Vasomotor Symptoms (Hot Flashes)

NCT03192176

ISN: ESN364_HF_205
IND number: 130277

Ogeda S.A.

Table of Contents

I. LIST OF ABBREVIATIONS AND KEY TERMS

List of Abbreviations

Abbreviations	Description of abbreviations
AE	Adverse Event
ALT	Alanine aminotransferase
ANCOVA	Analysis of covariance
ASCM	Analysis Set Classification Meeting
AST	Aspartate aminotransferase
ATC	Anatomical Therapeutic Chemical
AUC	Area under the curve
BALP	Bone alkaline phosphatase
BID	Bis in die; twice daily
BMC	Bone mineral content
BMD	Bone mineral density
BMI	Body mass index
BoT	Burden of Toxicity
CI	Confidence interval
CRF	Case Report Form
CRO	Contract Research Organization
CS	Classification Specifications
CSR	Clinical Study Report
C-SSRS	Columbia Suicide Severity Rating Scale
CTX	Carboxy-terminal telopeptide of type I collagen
CV	Coefficient of variation
DILI	Drug-induced liver injury
DXA	Dual-energy x-ray absorptiometry
E2	Estradiol
ECG	Electrocardiogram
eCRF	Electronic case report form
ePRO	Electronic Patient-Reported Outcome
ET	Early termination
FAS	Full Analysis Set
FDA	Food and Drug Administration
FSH	Follicle-stimulating hormone
FSI	First subject in
GCS	Greene Climacteric Scale
H	High
HBsAG	Hepatitis B virus surface antigen
HCV	Hepatitis C virus
HFRDIS	Hot Flash Related Daily Interference Scale
HIV	Human immunodeficiency virus
ICH	International Conference on Harmonization
IRT	Interactive Response Technology
L	Low
LH	Luteinizing hormone
LLN	Lower limit of normal
LSEQ	Leeds Sleep Evaluation Questionnaire

Abbreviations	Description of abbreviations
MedDRA	Medical Dictionary for Regulatory Activities
MENQoL	Menopause-Specific Quality of Life
MCP-Mod	Multiple Comparison Procedure – Modeling
mg	Milligram(s)
N	Normal
NIH	National Institutes of Health
P1NP	Procollagen type 1 amino-terminal propeptide
PD	Pharmacodynamic(s)
PGX	Pharmacogenomic(s)
PK	Pharmacokinetic(s)
PKAS	Pharmacokinetics Analysis Set
PPS	Per Protocol Analysis Set
PPS4	Per Protocol Analysis Set – Week 4
PPS12	Per Protocol Analysis Set – Week 12
PT	Preferred Term
QD	Quaque die; once daily
QT	Q-T Interval
QTcB	Corrected Q-T Interval (Bazett)
QTcF	Corrected Q-T Interval (Fridericia)
SA	Smooth average
SAE	Serious adverse event
SAF	Safety Analysis Set
SAP	Statistical Analysis Plan
SD	Standard deviation
SHBG	Sex hormone-binding globulin
SOC	System Organ Class
TEAE	Treatment-emergent adverse event
TLF	Tables, listings and figures
ULN	Upper limit of normal
USA	United States of America
WHO-DD	World Health Organization Drug Dictionary

List of Key Terms

Terms	Definition of terms
Baseline	A variable that represents the efficacy and safety measurements and disease characteristics prior to the intervention with study medication. It can be a single measurement or a derived parameter (such as average frequency or severity of vasomotor symptom from 7 non-missing days prior to Day 1).
Endpoint	A variable that pertains to the trial objectives.
Variable	Any quantity that varies; any attribute, phenomenon or event that can have different qualitative or quantitative values.
SAS ®	A statistical software package used for the analysis of data

1 INTRODUCTION

This Statistical Analysis Plan (SAP) contains a more technical and detailed elaboration of the principal features of the analysis described in the protocol, and includes detailed procedures for executing the statistical analysis of the primary and secondary endpoints and other data. This SAP is written based on Protocol Amendment 2: 05 March 2018.

The SAP will be finalized and signed prior to database hard lock and study unblinding.

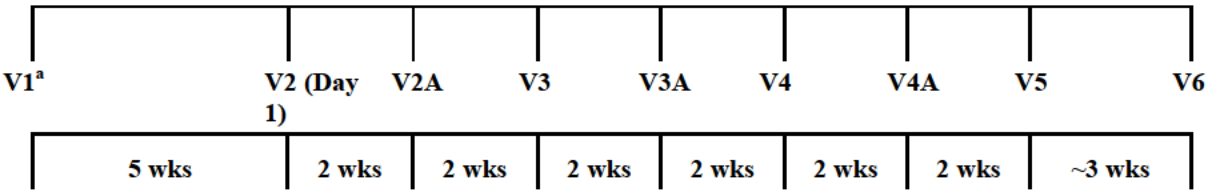
This statistical analysis is coordinated by the responsible biostatistician of Medpace, Inc. Any changes from the analyses planned in the SAP will be justified in the Clinical Study Report (CSR).

Prior to database hard lock, a final review of data and TLFs meeting will be held to allow a review of the clinical trial data and to verify the data that will be used for analysis set classification. If required, consequences for the statistical analysis will be discussed and documented. The classifications will be finalized prior to hard lock.

2 FLOW CHART AND VISIT SCHEDULE

2.1 Flow Chart

Screening	Randomization (1:1)	
	ESN364 15 mg BID ($N_{planned}$=44)	Follow-up
	ESN364 30 mg BID ($N_{planned}$=44)	Follow-up
	ESN364 60 mg BID ($N_{planned}$=44)	Follow-up
	ESN364 90 mg BID ($N_{planned}$=44)	Follow-up
	ESN364 30 mg QD + placebo QD ($N_{planned}$=44)	Follow-up
	ESN364 60 mg QD + placebo QD ($N_{planned}$=44)	Follow-up
	ESN364 120 mg QD + placebo QD ($N_{planned}$=44)	Follow-up
	Placebo BID ($N_{planned}$=44)	Follow-up



a. Screening is to be performed up to 35 days prior to randomization, with a minimum of 7 days to allow for baseline data collection of vasomotor symptom frequency and severity.

BID = twice daily; QD = once daily; V = visit; wks = weeks.

2.2 Schedule of Assessments

Assessments	**Screening Visit**	**Treatment Period**							**Follow-Up Visit**
Study Visit	**Visit 1**	**Visit 2**	**Visit 2A**	**Visit 3**	**Visit 3A**	**Visit 4**	**Visit 4A**	**Visit 5/ET**	**Visit 6**
Time of Visit	**Days -35 to -1[0,b]**	**Day 1[0]**	**Week 2[c]**	**Week 4[0]**	**Week 6[c]**	**Week 8[0]**	**Week 10[c]**	**Week 12[0]**	**Week 15[0,d]**
Informed consent[0]	X								
Informed consent PGX[0]	X								
Inclusion/exclusion criteria	X	X							
Medical history/concomitant diseases	X								
Screening mammogram[f]	X								
Demographic data[g]	X								
Physical examination[h]	X	X		X[i]		X[i]		X	X
Urine drug screen	X								
Urine pregnancy test	X								
Clinical laboratory[j] and urinalysis	X	X	X[k]	X	X[k]	X	X[k]	X	X
Alcohol breath test	X								
Vital signs[l]	X	X		X		X		X	X
12-lead ECG[0]	X	X		X		X		X	X
Pap smear[n]	X								
Transvaginal ultrasound	X							X	
Endometrial biopsy[o]	X							X	
Serology[p]	X								
Blood PD sample[q]		X		X		X		X	X
Blood PK sample[r]				X				X	
Blood PGX sample[s]								X	
Vasomotor symptom diary[t]	X	X		X		X		X	X
HFRDIS[u]		X		X		X		X	X
LSEQ[u]		X[v]		X[v]		X[v]		X[v]	X[v]
GCS[u]		X		X		X		X	X

Table continued on next page

Assessments	Screening Visit	Treatment Period							Follow-Up Visit
Study Visit	Visit 1	Visit 2	Visit 2A	Visit 3	Visit 3A	Visit 4	Visit 4A	Visit 5/ET	Visit 6
Time of Visit	Days -35 to -1[0,b]	Day 1[0]	Week 2[c]	Week 4[0]	Week 6[c]	Week 8[0]	Week 10[c]	Week 12[0]	Week 15[0,d]
MENQoL[u]		X		X		X		X	X
C-SSRS[u]		X						X	X
CCI	X							X	
Bone turnover markers[x]		X						X	X
ePRO training	X								
Randomization		X							
Dispense study drug[y]		X		X		X			
Study drug compliance and accountability[z]		X		X		X		X	
Concomitant medications and adverse events[aa]	X	X		X		X		X	X

a. Study visits should be conducted in the morning. Subjects should be fasted (defined as nothing by mouth except for water [up to 1 hour before study drug intake] for at least 10 hours prior to study visit). All assessments will be completed predose, except for postdose blood sampling for PK and PD assessments at Week 4 (Visit 3). Administration of questionnaires will be completed at the respective visits prior to any invasive procedures. Day 1 (Visit 2) should be planned to allow for the first intake of study drug at the study site between 7:00 AM and 10:00 AM. Following Day 1 (Visit 2), each subsequent visit should be planned at the same time of day (morning) for each subject. Following randomization to study drug, subjects will return to the study site for visits and procedures to occur within ±3 days of the scheduled time. Unscheduled visits can be planned outside the scheduled visits (see Section 6.6).

b. The screening visit is to occur within 35 days of randomization (Day 1 [Visit 2]), with a minimum of 7 days to allow for baseline data collection of vasomotor symptom frequency and severity assessments. Subjects will receive an electronic diary in which to record daily vasomotor symptoms during the screening period. Subjects must have ≥7 consecutive days of vasomotor symptom recordings to participate in the study. Subjects may be rescreened 1 time upon approval of the Medical Monitor. The following assessments do not need to be repeated at the rescreen provided they still fall within the acceptable time window: transvaginal ultrasound, endometrial biopsy, mammogram, CCI ECG, and Pap smear.

c. Subjects may schedule Week 2 (Visit 2A), Week 6 (Visit 3A), and Week 10 (Visit 4A) assessments at their convenience within the visit window (±3 days). Fasting is not required for these visits.

d. The follow-up visit will occur approximately 3 weeks following the last dose of study drug. For subjects requiring a repeat biopsy, the follow-up visit will occur 4 weeks after Week 12 (Visit 5).

e. Signed informed consent will be collected for all subjects before any study-related procedures are done. A separate signature will be collected for PGX sampling. Subjects who do not consent to PGX sampling are not excluded from participating in the study.

f. Only in the event the subject does not have a normal/negative or no clinically significant findings mammogram from previous 9 months on record.

Footnotes continued on next page

g. Includes age, race, sex, and smoking status (smoker/non-smoker).

h. Includes height (at the screening visit only), weight and waist circumference. A bimanual clinical pelvic and clinical breast examination will be performed at the screening visit. A bimanual clinical pelvic examination can be performed at any time in the study where clinically indicated.

i. Weight and waist circumference only.

j. Includes biochemistry, coagulation (at the screening visit and Week 12 [Visit 5]/ET only), and hematology panel. Blood samples for clinical laboratory tests should be taken in a fasted state (defined as nothing by mouth except water for 10 hours), except for Week 2 (Visit 2A), Week 6 (Visit 3A), and Week 10 (Visit 4A).

k. Includes biochemistry and hematology, only.

l. Includes oral/tympanic temperature, sitting blood pressure, and pulse rate (supine after 5 minutes of rest).

m. All 12-lead ECGs will be captured in triplicate (ie, 3 separate 12-lead ECGs at least 1 to 2 minutes apart within a 5-minute window). The subject should rest in supine position for at least 10 minutes prior to the first ECG.

n. Only in the event the subject does not have a normal/negative or no clinically significant findings Pap smear from previous 9 months on record.

o. Subjects will undergo endometrial biopsy at Screening (if endometrial thickness is ≥4 mm), at Week 12 (end-of-treatment)-(all subjects), at the ET Visit for subjects who are withdrawn from the study prior to completion (if study drug exposure is ≥10 weeks), and in all cases of uterine bleeding. If the uterine biopsy is abnormal at Week 12 (end-of-treatment), subjects will have a repeat biopsy 4 weeks later, if clinically indicated.

p. For HBsAG, anti-HCV antibodies, and anti-HIV antibodies.

q. Samples will be taken predose at Day 1 (Visit 2), Week 8 (Visit 4), and Week 12 (Visit 5)/ET, as well as predose and 3 hours postdose at Week 4 (Visit 3). The Week 4 (Visit 3) 3 hours postdose PD sample should be shifted to a later date in cases where the subject cannot accommodate the PK sampling schedule at that visit. The PD sample should be taken at the same time as the 3-hour postdose PK sample (see Section 8.2). A PD sample will be taken at the follow-up visit (Week 15 [Visit 6]) as well. Markers include LH, FSH, E2, and SHBG.

r. Samples to be taken Week 4 (Visit 3) predose and 1 h (±30 min), 3 h (±30 min), 5 h (-1 h/+30 min), and 7 h (-1 h/+2 h) postdose as well as predose Week 12 (Visit 5). The indicated time windows for Week 4 (Visit 3) PK sampling will allow for flexibility.

s. While scheduled for Week 12 (Visit 5), the PGX sample can be taken at any time during the study following signed informed consent and enrollment into the study.

t. The vasomotor symptom diary will be kept by subjects. Subjects will record the vasomotor symptom frequency and severity of each vasomotor symptom in the ePRO diary on an ad hoc basis, but at a minimum twice daily (morning and evening), from the screening visit (Days -35 to -1 [Visit 1]) through the follow-up visit (Week 15 [Visit 6]). Night sweats should be recorded no later than in the morning upon awakening to start a new day.

u. Paper-based, administered at the study site prior to any invasive procedures.

v. A baseline LSEQ assessment will be collected at Day 1 (Visit 2). A post-baseline LSEQ assessment will be collected for all other indicated visits.

w. CCI [redacted]

x. Includes BALP, P1NP, and CTX.

Footnotes continued on next page

y. Subjects will be assigned 2 bottles of study drug as a kit (1 bottle labeled “morning” and 1 bottle labeled “evening”) at Day 1 (Visit 2), Week 4 (Visit 3), and Week 8 (Visit 4). Each bottle will contain 34 capsules of ESN364 or placebo. Study drug intake will be done with a glass of room temperature tap water. The first intake of study drug will take place at the study site on Day 1 (Visit 2) between 7:00 AM and 10:00 AM, under the supervision of the study staff. On study visit days (Day 1 [Visit 2], Week 4 [Visit 3], and Week 8 [Visit 4], only), the morning dose of study drug will be taken at the study site, under the supervision of the study staff, after collection of predose blood samples. Subjects will be instructed to arrive at the study site fasted (nothing by mouth except water [up to 1 hour prior to study drug intake] for at least 10 hours). On all other days throughout the treatment period, subjects will be instructed to take their morning dose of study drug at home, around the same time of the day spaced as near to 12 hours apart as possible (preferably between 7:00 AM and 10:00 AM and 7:00 PM and 10:00 PM). Intake of study drug should be preferably 1 hour before or 2 hours after meals. Subjects need to record all home study drug intake in the ePRO diary.

z. Includes compliance of study drug intake and ePRO completion. Subjects will be asked to return all unused study drug. Compliance of study drug intake will be assessed by counting returned study drug in addition to reviewing ePRO entries of study drug intake. Any discrepancies between returned study drug number and dosing in the ePRO diary will be discussed with the subject for whom a discrepancy was seen and recorded in the source documents and the eCRF.

aa. Adverse events and intake of concomitant medication(s) will be monitored continuously from informed consent until the last study-related activity.

BALP = bone alkaline phosphatase; C-SSRS = Columbia Suicide Severity Rating Scale; CTX = carboxy-terminal telopeptide of type I collagen; CCI [redacted] E2 = estradiol; ECG = electrocardiogram; eCRF = electronic Case Report Form; ET = early termination; ePRO = electronic Patient-Reported Outcome; FSH = follicle-stimulating hormone; GCS = Greene Climacteric Scale; HBsAG = hepatitis B virus surface antigen; HCV = hepatitis C virus; HFRDIS = Hot Flash Related Daily Interference Scale; HIV = human immunodeficiency virus; LH = luteinizing hormone; LSEQ = Leeds Sleep Evaluation Questionnaire; MENQoL = Menopause-Specific Quality of Life; P1NP = procollagen type 1 amino-terminal propeptide; PD = pharmacodynamic; PGX = pharmacogenomic; PK = pharmacokinetic; SHBG = sex hormone-binding globulin.

3 STUDY OBJECTIVE(S) AND DESIGN

3.1 Study Objective(s)

3.1.1 Primary Objective

The primary objective of this study is to evaluate the effect of different doses and dosing regimens of ESN364 on frequency and severity of vasomotor symptoms (hot flashes).

3.1.2 Secondary Objectives

The secondary objectives of this study are the following:

- To evaluate the effect of different doses and dosing regimens of ESN364 on the frequency, severity, and hot flash score of mild, moderate, and severe vasomotor symptoms;
- To evaluate the effect of different doses and dosing regimens of ESN364 on responder rates using a variety of responder definitions;
- To evaluate the effect of different doses and dosing regimens of ESN364 on patient-reported outcomes;
- To evaluate the effect of different doses and dosing regimens of ESN364 on PD markers (hormone and bone markers); and
- To evaluate the effect of different doses and dosing regimens of ESN364 on safety and tolerability.

3.1.3 Exploratory Objectives

The exploratory objectives of this study are the following:

- CCI [redacted]
- To evaluate the PK plasma concentrations of ESN364 and metabolite CCI [redacted]

3.2 Study Design

This is a 12-week, randomized, double-blind, placebo-controlled, dose-ranging, parallel-group, multicenter study to assess the efficacy of ESN364 in postmenopausal women suffering from vasomotor symptoms (hot flashes).

This study will consist of a screening period (Days -35 to -1, including the screening visit [Visit 1] and a minimum 7-day collection of baseline vasomotor symptom frequency and severity assessments), a 12-week treatment period (Day 1 [Visit 2] to Week 12 [Visit 5], including safety visits at Week 2 [Visit 2A], Week 6 [Visit 3A], and Week 10 [Visit 4A]), and a follow-up visit (Week 15 [Visit 6]) approximately 3 weeks after the last dose of study drug, for a total of 9 visits.

The study will be performed on an ambulatory basis.

The screening visit (Visit 1) will occur up to 35 days prior to randomization. Eligibility will be assessed via a physical examination, clinical laboratory testing, vital signs, ECG, Pap

smear, mammography, and endometrial biopsy. Subjects will receive an electronic diary in which to record daily vasomotor symptoms during the duration of the screening period. Subjects must have ≥7 consecutive days of vasomotor symptom recordings to participate in the study. Subjects are encouraged to continue recording for the duration of the whole screening period. The electronic diary will be reviewed by study site staff on Day 1 (Visit 2) to confirm study eligibility. Subjects may be rescreened 1 time upon approval of the medical monitor. During the treatment period, subjects will return to the study site every 2 weeks for assessments. The follow-up visit will occur approximately 3 weeks following the last dose of study drug.

Approximately 352 subjects will be enrolled in the study. Subjects will be randomized to 1 of 8 treatment groups (i.e., 44 subjects in each arm).

Subjects and study personnel will be blinded to the treatment (placebo or ESN364), dosage (30 mg, 60 mg, etc.), and dosing regimen (QD or twice daily [BID]).

3.3 Randomization

In total, approximately 352 subjects will be randomized to 1 of 8 treatment groups in an equal ratio.

Randomization will be based on a computer-generated randomization schedule prepared by an unblinded biostatistician from Medpace Biostatistics using SAS® software (SAS Institute Inc., Cary, North Carolina, USA) prior to the start of the study. The randomization will be balanced using randomly permuted blocks across the treatment arms. The randomization list will be retained by Medpace Randomization and Study Product Management until the end of the study (database lock). A copy will be sent in a sealed envelope to the bioanalytical laboratory responsible for plasma drug and serum hormone determination before the start of the study as applicable. Because of a planned early timepoint analysis (see Section 7.9) and because PK/PD modeling will start prior to database hard-lock, specific assigned Sponsor staff involved in these activities (or delegates, if outsourced to an external vendor) will be unblinded after recruitment has been completed. Details of the unblinding will be documented in a separate charter to ensure the direct study team remains blinded and that unblinded results are restricted to a very small group independent of the study team.

Subjects will be assigned a randomization number from the Interactive Response Technology (IRT) system upon enrollment. The randomization list will be uploaded into the IRT system. Based on this randomization list, the study drug will be packaged and labeled. Medication numbers will be preprinted on the study drug labels. The study team, including those involved in data management, clinical, medical, or statistical review of the study data, will also be blinded during the study as they will not have access to the randomization list. Medpace, the Contract Research Organization (CRO) performing data management and statistical activities, will receive a copy of the randomization list after database lock. The study drug administered to a subject can be identified by the IRT system. The IRT system will allow rapid access to the treatment allocation codes when relevant for site or CRO personnel.

4 SAMPLE SIZE

A total of 352 subjects are planned to be randomized, 44 subjects in each treatment arm.

In the Phase 2a study, the observed least-squares mean difference between ESN364 90 mg BID (highest dose of this Phase 2b study) and placebo in change from baseline to Week 12 in mean daily frequency of moderate to severe vasomotor symptoms was -5.0 (95% CI: - 6.8, -3.3) with similar results at Week 4. For any given pairwise comparison for a 2-sample t-test at a 2-sided 5% alpha, 40 subjects provides the following power to detect the following effect sizes assuming an SD of 5:

Assumed treatment difference in mean daily frequency	Power for pairwise test
3.3	83%
3.5	87%
4	94%
4.5	97%
5	>99%

For change from baseline to Week 12 in mean severity of moderate to severe vasomotor symptoms, the observed mean treatment difference in the Phase 2a study was -1.12 (95% CI: -1.5, -0.74) with similar results at Week 4. For any given pairwise comparison for a 2-sample t-test at a 2-sided 5% alpha, 40 subjects provides the following power to detect the following effect sizes assuming an SD of 1:

Assumed treatment difference in mean severity	Power for pairwise test
0.64	80%
0.75	91%
1	>99%

Note that the combined power for testing all 4 co-primary endpoints will be lower than the power for each considered individually. Assuming approximately 10% of subjects discontinue prematurely, the number of 40 subjects is increased to 44 subjects per arm.

5 ANALYSIS SETS

In accordance with International Conference on Harmonization (ICH) recommendations in guidelines E3 and E9, the following analysis sets will be used for the analyses.

Detailed criteria for analysis sets will be laid out in Classification Specifications (CS) and the allocation of subjects to analysis sets will be determined prior to database hard lock.

All primary and secondary efficacy endpoints will be analyzed using the Full Analysis Set (FAS). The Per Protocol Analysis Set (PPS) will be used only for the analysis of selected endpoints to examine the robustness of the primary analysis. Safety and tolerability will be analyzed using the Safety Analysis Set (SAF).

5.1 Safety Analysis Set (SAF)

The SAF comprises all randomized subjects who received at least one dose of study drug. The SAF will analyze patients as treated. In case a subject erroneously received a treatment different from her randomized treatment, the subject will be assigned to the treatment group

that the subject received as first dose. The SAF will be used for summaries of demographic and baseline characteristics and all safety and tolerability related variables.

5.2 Full Analysis Set (FAS)

The FAS comprises the subset of subjects from the SAF who received at least one dose of study drug, had a baseline and at least one post-baseline efficacy evaluation. The FAS will analyze patients as randomized, irrespective of the actual treatment received. The FAS will be used for summaries and primary analyses of efficacy data, as well as selected demographic and baseline characteristics.

5.3 Per Protocol Analysis Sets (PPS)

The PPS comprises the subset of subjects from the FAS who are treated according to the protocol without any major deviations. A full detail of inclusions and exclusions from Per Protocol Analysis Set - Week 4 (PPS4) and Per Protocol Analysis Set - Week 12 (PPS12) will be in the classifications specifications and finalized before unblinding.

5.3.1 Reasons for Exclusion From PPS4

The following reasons may lead to subject's exclusion from PPS4:

- No measurement of the primary efficacy endpoint available at Week 4.
- <85% interactive diary compliance during the 4 week treatment period.
- Treatment compliance less than or equal to 85% between randomization and Week 4.

5.3.2 Reasons for Exclusion From PPS12

The following reasons may lead to subject's exclusion from PPS12:

- No measurement of the primary efficacy endpoint available at Week 12.
- <85% interactive diary compliance during the 12 week treatment period.
- Treatment compliance less than or equal to 85% between randomization and Week 12.

5.4 Pharmacokinetics Analysis Set (PKAS)

The PKAS comprises the subset of subjects who receive study drug and provide at least one sample for the measurement of drug concentrations. Exclusion of subjects from the PKAS will be considered by the pharmacokineticist on a case-by-case basis.

6 ANALYSIS VARIABLES

6.1 Efficacy Endpoints

6.1.1 Co-Primary Efficacy Endpoints

The primary efficacy objectives will require the evaluation of the effect of ESN364 on the following 4 co-primary endpoints:

- Mean change in the frequency of moderate to severe vasomotor symptoms from baseline to Week 4;

- Mean change in the frequency of moderate to severe vasomotor symptoms from baseline to Week 12;
- Mean change in the severity of moderate to severe vasomotor symptoms from baseline to Week 4; and
- Mean change in the severity of moderate to severe vasomotor symptoms from baseline to Week 12.

Subjects will record the number of vasomotor symptoms and severity of each vasomotor symptom via the ePRO diary from the screening visit (Days -35 to -1 [Visit 1]) through the follow-up visit (Week 15 [Visit 6]).

Night sweats (vasomotor symptoms that disrupt the subject's night sleep) should be recorded no later than in the morning upon awakening to start a new day. Subjects should take note of the time of the night sweat and record it on the appropriate calendar day (i.e., before or after midnight).

The 24 hour vasomotor symptoms runs from 08:00 AM to 07:59 AM (next day); the 12 hour day time vasomotor symptoms is between 08:00 AM and 07:59 PM; the 12 hour night time vasomotor symptoms is between 08:00 PM and 07:59 AM (next day) regardless of whether the patient was awake or woke up because of the event.

Frequency of moderate to severe vasomotor symptoms is the number of moderate to severe vasomotor symptoms per 24h (or 12h day time or 12h night time).

The severity of vasomotor symptoms is defined clinically as follows (according to the FDA and European Medicines Agency Guidances for Industry[1,2] and the National Institutes of Health [NIH] Hot Flash Workshop[3]):

- Mild: Sensation of heat without sweating/dampness. If at night, subject does not wake up but later notices damp sheets or clothing.
- Moderate: Sensation of heat with sweating/dampness, but able to continue activity. If at night, subject wakes up because she is feeling hot and/or is sweating, but no action is necessary other than rearranging the bed sheets.
- Severe: Sensation of intense heat with sweating, causing disruption of activity. If at night, subject wakes up hot and is sweating and needs to take action (e.g., remove layers of clothes, open the window, or get out of bed).

The severity of moderate and severe vasomotor symptoms per 24h (or 12h day time or 12h night time):

> [(number of moderate vasomotor symptoms x 2) +
> (number of severe vasomotor symptoms x 3)] / number of moderate/severe vasomotor symptoms.

Severity is zero for subjects who have no moderate or severe vasomotor symptoms.

A daily frequency and severity per week will be derived by taking the mean of the data over 7 days. A daily average per week will be derived if any information for 3 or more days was reported (morning dose, evening dose, or vasomotor symptom).

Because baseline data is preferably the most recent data prior to randomization and dosing, baseline for this study is the mean value of vasomotor symptoms calculated over the last 7 calendar days with non-missing data prior to Day 1 (Visit 2). As a consequence, the baseline might be less than the required 50 moderate/severe hot flashes per week required for study entry, as these can occur anywhere during the 35 day screening period provided that they are the last 7 days with non-missing data prior to Day 1 (Visit 2).

See Section 7.10.6 for study week and endpoint derivations.

6.1.2 Secondary Efficacy Endpoints

The secondary efficacy endpoints include the effect of ESN364 on the following:

6.1.2.1 Vasomotor Symptom Endpoints:

- Mean change in the frequency of mild, moderate, and severe vasomotor symptoms from baseline to each study week;
- Mean change in the frequency of moderate and severe vasomotor symptoms per 24h from baseline to each study week;
- Mean change in the severity of mild, moderate, and severe vasomotor symptoms from baseline to each study week;
- Mean change in the severity of moderate and severe vasomotor symptoms from baseline to each study week;
- Mean change in the hot flash score of mild, moderate, and severe vasomotor symptoms from baseline to each study week;
- Mean change in the hot flash score of moderate and severe vasomotor symptoms from baseline to each study week;
- Mean percent reduction of mild, moderate, and severe vasomotor symptoms from baseline to each study week; and
- Mean percent reduction of moderate and severe vasomotor symptoms from baseline to each study week.

Frequency of mild, moderate and severe vasomotor symptoms is the number of mild, moderate and severe vasomotor symptoms per 24h (or 12h day time or 12h night time).

Severity of mild, moderate and severe vasomotor symptoms per 24h (or 12h day time or 12h night time):

> [(number of mild vasomotor symptoms x 1) +
> (number of moderate vasomotor symptoms x 2) +
> (number of severe vasomotor symptoms x 3)] / number of mild/moderate/severe vasomotor symptoms

Severity is zero for subjects who have no hot flashes.

Hot flash score per 24h (or 12 h day time or 12 h night time): The hot flash score per 24h (or 12 h day time or 12 h night time) of vasomotor symptoms (mild, moderate, and severe) is calculated as follows:

(number of mild vasomotor symptoms x 1) +
(number of moderate vasomotor symptoms x 2) +
(number of severe vasomotor symptoms x 3).

The hot flash score per 24h (or 12 h day time or 12 h night time) of moderate and severe vasomotor symptoms is calculated as follows:

(number of moderate vasomotor symptoms x 2) +
(number of severe vasomotor symptoms x 3).

A weekly frequency, severity, or hot flash score will be derived similar to the co-primary efficacy endpoints. See Section 7.10.6 for study week and endpoint derivations.

6.1.2.2 Responder Endpoints:

- Mean percent reduction of 50%, 70%, 90%, and 100% of mild, moderate, and severe vasomotor symptoms from baseline to each study week;
- Mean percent reduction of 50%, 70%, 90%, and 100% of moderate and severe vasomotor symptoms from baseline to each study week;
- Absolute reduction of 2, 3, 4, and 5 in mean number of mild, moderate, and severe vasomotor symptoms per 24h from baseline to each study week; and
- Absolute reduction of 2, 3, 4, and 5 in mean number of moderate and severe vasomotor symptoms per 24h from baseline to each study week.

6.1.2.3 Patient-Reported Outcome Endpoints:

- Change in Hot Flash Related Daily Interference Scale (HFRDIS) from baseline to Weeks 4, 8, 12, and 15;
- Change in Leeds Sleep Evaluation Questionnaire (LSEQ) from baseline to Weeks 4, 8, 12, and 15;
- Change in Greene Climacteric Scale (GCS) from baseline to Weeks 4, 8, 12, and 15; and
- Change in Menopause-Specific Quality of Life (MENQoL) from baseline to Weeks 4, 8, 12, and 15.

HFRDIS: The HFRDIS is a 10-item scale that measures a woman's perceptions of the degree to which vasomotor symptoms interfere with 9 daily life activities (work, social activities, leisure, sleep, mood, concentration, relations with others, sexuality, and enjoying life); the 10th item measures interference with overall quality of life.[5] This scale was modeled after items on the Brief Pain Inventory[6] and Brief Fatigue Inventory,[7] which assess the extent to which pain or fatigue interfere with daily life. Subjects will be asked to rate the extent to which vasomotor symptoms have interfered with each item during the previous 2-week time interval using a 0 (do not interfere) to 10 (completely interfere) scale. Recent structural equation modeling suggests the HFRDIS is a unidimensional scale best represented by an overall mean score calculated as the sum of items/10.[8] If any of the 10 items have missing values, the overall mean score will not be calculated.

LSEQ: The LSEQ is a 10-item self-rated questionnaire that assesses a subject's aspects of sleep and early morning behavior. The questions are grouped into 4 chronological domains: ease of getting to sleep (items 1 to 3), perceived quality of sleep (items 4 to 5), ease of

awaking from sleep (items 6 to 7) and integrity of early morning behavior following wakefulness (items 8 to 10).[9] The LSEQ is a visual analogue scale that requires respondents to place marks on a group of 10 cm lines representing the changes they have experienced in a variety of symptoms compared to usual (baseline version) versus since the beginning of treatment (post-baseline version). Lines extend between extremes like "more difficult than usual" and "easier than usual." Responses are measured using a 100 mm scale and are averaged to provide a score for each domain. Note the questionnaire was reformatted for data collection to ensure precise measurements. Visit 2 data not collected on the baseline version of the LSEQ will be excluded from the analysis and listings. If any of the responses in a domain have missing values, the score for the domain will not be calculated.

GCS: The GCS is a 21-item scale that provides a brief but comprehensive and valid measure of climacteric symptomatology.[10] Each item is rated by the subject according to its severity using a 4-point rating scale from 0 (none) to 3 (severe). The first 20 items of the scale combine into 3 main independent symptom measures by summing up the individual item scores:

- Psychological symptoms (items 1 to 11; score 0 to 33),
- Physical symptoms (items 12 to 18; score 0 to 21), and
- Vasomotor symptoms (items 19 to 20; score 0 to 6).

Item 21 is a probe for sexual dysfunction. The total score can range from 0 to 63. Higher scores indicate worse symptoms. If any of the contributing items have missing values, scores for psychological symptoms, physical symptoms, or vasomotor symptoms will not be calculated. If any of the scores for psychological symptoms, physical symptoms, or vasomotor symptoms are missing, the total score will not be calculated.

MENQoL: The MENQoL is a self-administered questionnaire and consists of a total of 29 items in a Likert-scale format. Each item assesses the impact of 1 of 4 domains of menopausal symptoms, as experienced over the last month:

- Vasomotor (items 1 to 3),
- Psychosocial (items 4 to 10),
- Physical (items 11 to 26), and
- Sexual (items 27 to 29).

Items pertaining to a specific symptom are rated as present or not present, and if present, how bothersome on a 0 (not bothersome) to 6 (extremely bothersome) scale.[11,12] Non-endorsement of an item is scored a "1" and endorsement a "2" plus the number of the particular rating, so that the possible score of any item ranges from 1 to 8.[12] Mean scores are computed for each domain by dividing the sum of the domain's items by the number of items within that domain.[13] The overall mean score is the mean of the domain mean scores. If any of the contributing items have missing values, the mean score for the domain will not be calculated. If any of the domain mean scores are missing, the overall mean score will not be calculated.

For all patient reported outcome endpoints, baseline is the value at Day 1 (Visit 2). Questionnaires are completed at site prior to any invasive procedures are completed.

6.1.2.4 Pharmacodynamics Markers

- Change from baseline to Week 12 in plasma concentrations of luteinizing hormone (LH), follicle-stimulating hormone (FSH), estradiol (E2), and sex hormone-binding globulin (SHBG).

6.1.3 Exploratory Efficacy Endpoints

The exploratory endpoints include the effect of ESN364 on the following:

- CCI [redacted]
- Level of PK plasma concentrations of ESN364 and metabolite CCI [redacted] at pre-specified timepoints;
- Change in weight and waist circumference from baseline to Weeks 4, 8, 12, and 15;
- Mean change in the frequency of moderate to severe 12h day time vasomotor symptoms from baseline to Week 4;
- Mean change in the frequency of moderate to severe 12h day time vasomotor symptoms from baseline to Week 12;
- Mean change in the severity of moderate to severe 12h day time vasomotor symptoms from baseline to Week 4;
- Mean change in the severity of moderate to severe 12h day time vasomotor symptoms from baseline to Week 12;
- Mean change in the frequency of moderate to severe 12h night time from baseline to Week 4;
- Mean change in the frequency of moderate to severe 12h night time from baseline to Week 12;
- Mean change in the severity of moderate to severe 12h night time from baseline to Week 4; and
- Mean change in the severity of moderate to severe 12h night time from baseline to Week 12.

6.2 Safety Variables

Safety variables will include the following:

- Incidence and severity of treatment-emergent adverse events (TEAEs);
- Endometrial health assessment (transvaginal ultrasound ± endometrial biopsy);
- Vital signs (sitting systolic and diastolic blood pressure and pulse rate);
- Laboratory tests (hematology, biochemistry, urinalysis, and coagulation);
- 12-lead electrocardiogram (ECG) parameters;
- Plasma bone density marker concentrations;

- Physical examinations; and
- Columbia Suicide Severity Rating Scale (C-SSRS).

A TEAE is defined as an adverse event observed after starting administration of study drug through 7 days after the last dose of study drug. If the adverse event occurs on the day of the first dose and the onset time is on or after the time of the first dose of study drug or the onset time is missing, then the adverse event will be considered treatment emergent. If the adverse event occurs on the day of first dose and the onset time is before the time of the first dose of study drug, then the adverse event will not be considered treatment emergent. If a subject experiences an event both during the pre-investigational period that continues into the investigational period, the event will be considered as TEAE only if it has worsened in severity (i.e. it is reported with a new start date).

A drug-related TEAE is defined as any TEAE related to study treatment as assessed by the investigator or with missing assessment of the causal relationship.

Other safety endpoints will be considered treatment emergent if observed after starting administration of study drug through 1 day after the last dose of study drug.

6.3 Pharmacokinetic Variables

The plasma concentration data may be used in a meta-population pharmacokinetic analysis. The results of the population PK analysis for ESN364 and its active metabolite will be described in separate report.

6.4 Other Variables

Duration of treatment exposure

Duration of exposure to treatment will be calculated in days, using the following formula:

Days of exposure = date of last dose – date of first dose + 1

Study drug compliance

Study drug compliance will be calculated using the following formula:

$$\text{Study drug compliance} = 100 \times \frac{\text{actual capsules consumed}}{\text{expected capsules consumed}}$$

where

Actual capsules consumed = total capsules dispensed – total capsules returned

Expected capsules consumed = (last dose date – first dose date +1) x 2 capsules per day

Study drug compliance of morning dose

Study drug compliance of morning dose will be calculated using the following formula:

$$\text{Compliance of morning dose} = 100 \times \frac{\text{actual capsules consumed}}{\text{expected capsules consumed}}$$

where

Actual capsules consumed = morning capsules dispensed – morning capsules returned

Expected capsules consumed = (last dose date – first dose date +1) capsules per day

Study drug compliance of evening dose

Study drug compliance of evening dose will be calculated using the following formula:

$$\text{Compliance of evening dose} = 100 \times \frac{\text{actual capsules consumed}}{\text{expected capsules consumed}}$$

where

Actual capsules consumed = evening capsules dispensed – evening capsules returned

Expected capsules consumed = (last dose date – first dose date +1) capsules per day

Diary interaction compliance during the treatment period

Diary compliance during the treatment period will be calculated using the following formula:

$$\text{Diary interaction compliance} = 100 \times \frac{\text{actual diary days with data recorded}}{\text{expected diary days with data recorded}}$$

where

Actual diary days with data recorded = number of diary days with any morning or evening dose or hot flash data recorded from Visit 2 up to and including the last day of exposure to treatment

Expected diary days with data recorded = date of last dose – date of Visit 2 + 1

Previous and concomitant medication

Previous medication is defined as medication with at least one dose taken before the date of the first dose of study drug.

Concomitant medication is defined as medication with at least one dose taken between the date of first dose (inclusive) and the date of last dose (inclusive) of study drug.

7 STATISTICAL METHODOLOGY

7.1 General Considerations

For continuous variables, descriptive statistics will include the number of subjects (n), mean, standard deviation, minimum, median, maximum, 10%, 90%, and lower and upper quartiles. Frequencies and percentages will be displayed for categorical data. Percentages by categories will be based on the number of subjects with no missing data, i.e. will add up to 100%.

Summaries based on FAS and PPS (e.g. disposition, baseline and efficacy data) will be presented by planned treatment group, unless specifically stated otherwise. Safety analysis and other summaries based on SAF will be presented by actual treatment received.

All statistical comparisons will be conducted using two sided tests at the α=0.05 significance level unless specifically stated otherwise. All null hypotheses will be of no treatment difference, all alternative hypotheses will be two-sided, unless specifically stated otherwise.

All data processing, summarization, and analyses will be performed using SAS® Version 9.3 or higher on Windows®. Specifications for tables, figures, and data listing formats can be found in the TLF specifications.

7.2 Study Population

7.2.1 Disposition of Subjects

The following subject data will be presented:

- Number of subjects with informed consent, discontinued before randomization, randomized (overall only);
- Number and percentage of subjects randomized in each analysis set, by treatment group and overall;
- Number and percentage of subjects completed and discontinued the study, by primary reason for study discontinuation for randomized subjects and by treatment group;
- Number and percentage of subjects excluded from PPS4 and PPS12 by reason for exclusion defined in Section 5.3.1 and 5.3.2, by treatment group for FAS.

7.2.2 Protocol Deviations

Protocol deviations as defined in the Medpace Protocol Deviation Plan will be summarized for all randomized subjects. The number and percentage of subjects meeting any criteria will be summarized for each criterion and overall, by treatment group and total as well as by study site. Subjects deviating from a criterion more than once will be counted once for the corresponding criterion. Any subjects who have more than one protocol deviation will be counted once in the overall summary. A data listing will be provided by site and subject.

The protocol deviation criteria will be uniquely identified in the summary table and listing.

7.2.3 Demographic and Other Baseline Characteristics

Demographic and other baseline characteristics will be summarized by descriptive statistics.

Number and percentage of subjects randomized in each country and site will be presented by treatment group for the SAF.

Descriptive statistics for age, weight, body mass index (BMI) and height at study entry will be presented. Frequency tabulations for sex, ethnicity, and race will be presented. This will be done for the non-randomized subjects, as well as for the SAF and FAS, by treatment group.

Medical history is coded in MedDRA, and will be summarized by System Organ Class (SOC) and Preferred Term (PT) as well as by PT alone, by treatment group for the SAF.

Disease-specific medical history including spontaneous amenorrhea and bilateral oophorectomy will be summarized by treatment group for the SAF. The number and percentage of subjects with amenorrhea <6 months, >=6 - <12 months or >=12 months will be presented by treatment group for the SAF.

Social history (smoking habits) will be summarized by treatment group for the SAF.

7.2.4 Previous and Concomitant Medications

Previous medications are coded with WHO-DD, and will be summarized by anatomic main group (ATC 1st level), therapeutic subgroup (ATC 2nd level), chemical subgroup (ATC 4th level), and preferred WHO name by treatment group for the SAF.

As with previous medications, concomitant medications will be summarized for each treatment group by anatomic main group (ATC 1st level), therapeutic subgroup (ATC 2nd level) and chemical subgroup (ATC 4th level) and preferred WHO name for the SAF. Subjects taking the same medication multiple times will be counted once per medication and investigational period. A medication which can be classified into several chemical and/or therapeutic subgroups is presented in all chemical and therapeutic subgroups.

7.3 Study Drugs

7.3.1 Exposure

The following information on drug exposure will be presented for each treatment group for the SAF:

- Descriptive statistics for cumulative amount of the drug subject was exposed to and average daily dose; and
- Number and percent of subject with dose interruptions by treatment group.

Duration of exposure will be summarized in two ways.

- Descriptive statistics will be presented by treatment group.
- Exposure time will be categorized according to the following categories by treatment group:
    - less than 28 days
    - at least 28 days, less than 57 days
    - at least 57 days, less than 72 days
    - at least 72 days, less than 93 days
    - At least 93 days
    - Unknown.

Counts and percentages of subjects in each of these categories will be summarized for each treatment group for the SAF.

7.3.2 Treatment Compliance

Overall treatment compliance with the dosing schedule, compliance of morning dose, and compliance of evening dose will be examined for subjects in the SAF whose total study drug count and first and last days of treatment are known. Compliance will be calculated compared to the actual treatment period of dosing (first to last day of treatment), not to the planned treatment period.

Percent overall compliance, percent compliance of morning dose, and percent compliance of evening dose will be summarized in two ways for the SAF:

- Descriptive statistics will be presented by treatment group.
- Percent compliance will be categorized according to the following categories by treatment group:
    - Less than 50%,
    - at least 50%, less than or equal to 85%,
    - greater than 85%, less than or equal to 120%,
    - Over 120%, and
    - Unknown.

Overall treatment compliance, compliance of morning dose, and compliance of evening dose through Week 4 will be analyzed in a similar manner.

7.4 Analysis of Efficacy

7.4.1 Analysis of Primary Endpoints

7.4.1.1 Primary Analysis of Co-Primary Endpoints

The primary efficacy analysis will be based on the FAS.

The mean frequency and severity of moderate to severe vasomotor symptoms as well as change from baseline will be summarized by study week and for the last on treatment week. See Section 7.10.6 for study week and endpoint derivations. Weekly means will be plotted.

For each of the co-primary efficacy endpoints, an analysis of covariance (ANCOVA) model will be used with treatment group, pooled center, and smoking status (current vs. former/never) as factors, with baseline weight and baseline measurement as covariates. Pooled centers will be generated and documented prior to database lock. The algorithm for pooling centers can be found in Section 7.10.2. If the inclusion of pooled centers creates modeling problems, then dropping that factor will be considered.

Pairwise comparisons between the active doses and placebo will be calculated based on least-squares mean contrasts using a 2-tailed 95% confidence interval (CI).

For subjects in the efficacy analysis populations with missing primary efficacy endpoints, multiple imputation by fully conditional specification methods will be used. The multiple imputation model is specified in Section 7.10.1

Since the study design requires the comparison of 7 active dose groups with placebo for 4 co-primary efficacy endpoints, a two-tier closed testing procedure will be used to control the

family-wise error rate. A step-down testing procedure will be followed. The following testing order of doses will be employed:

a. Placebo versus 90 mg BID (daily dose 180 mg),

b. Placebo versus 60 mg BID (daily dose 120 mg),

c. Placebo versus 120 mg QD (daily dose 120 mg),

d. Placebo versus 30 mg BID (daily dose 60 mg),

e. Placebo versus 60 mg QD (daily dose 60 mg),

f. Placebo versus 15 mg BID (daily dose 30 mg), and

g. Placebo versus 30 mg QD (daily dose 30 mg).

The 4 co-primary efficacy endpoints will be compared at each dose level. If all 4 co-primary efficacy endpoints are statistically significantly better than placebo at a 2-sided alpha level of 0.05 for a given dose this dose level will be considered superior to placebo. A failure to prove the significance for one or more co-primary efficacy endpoints will lead to a failure to demonstrate superiority to placebo at this dose level, and the other dose levels after this one in the sequence will not be tested. In this testing sequence, if one comparison failed to provide the significance at this dose level, the comparisons at other dose levels following this sequence will not be tested.

7.4.1.2 Secondary Analysis of Co-Primary Endpoints

The primary analysis for each of the co-primary endpoints will be repeated using an ANCOVA model with treatment group as a factor and baseline measurement as a covariate.

A supportive analysis will be carried out for the co-primary efficacy endpoints based on the PPS to examine the impact of premature dropouts and/or major protocol deviations. The method used for this analysis will be identical to the primary analysis described in Section 7.4.1.1. The missing value imputation method for the early termination might not be applicable for the analysis on PPS since the analysis set only includes completed subjects. Week 4 analyses will be performed on PPS4, and the Week 12 analyses will be performed on PPS12.

An analysis of the primary efficacy variable by week will be conducted with mixed model for repeated measurements (MMRM) on FAS. This analysis will use a restricted maximum likelihood (REML)–based repeated-measures approach. The analyses will include the treatment group, week and smoking status (current vs. former/never) as factors and baseline measurement as a covariate, as well as an interaction of treatment by week and an interaction of baseline measurement by week. An unstructured covariance structure shared across treatment groups will be used to model the within-patient errors. The Kenward-Roger approximation will be used to estimate denominator degrees of freedom and adjust standard errors. The treatment difference will be estimated at all study weeks.

In addition, MCP-Mod analyses as described in the MCP-Mod SAP will be performed on co-primary endpoints.

7.4.2 Analysis of Secondary Endpoints

The analysis of secondary efficacy endpoints will be based on the FAS.

For each of the secondary vasomotor symptom and patient-reported outcome endpoints, an MMRM model as described in Section 7.4.1.2 will be used. The treatment difference will be estimated at all study weeks (week replaced by visit for patient-reported outcome endpoints).

For each of the secondary responder endpoints, logistic regression will be used for the analysis for each week and endpoint. The analyses will include the treatment group and smoking status (current vs. former/never) as factors and baseline measurement (mean frequency of vasomotor symptoms) as a covariate. The missing value will be imputed as a non-responder.

For these responder endpoints, the number of subjects who achieve the reduction will be summarized by study week with descriptive statistics for each treatment arm. For 50% and 90% reduction responders, Kaplan-Meier estimate will be used to estimate the time to response and the time to sustained response. Smooth average (SA) will be used for the time to response. SA value for day X-1, day X and Day X+1 would be assigned to day X. The time to response will be the first day with at least 50% (or 90%) response for both original numbers and SA number. The time to sustained response will be the first day with at least 50% (or 90%) response for both original numbers and SA number, and SA response from that day till the end of the treatment at least 50% (or 90%) response on each day. For the non-responders, it will be censored at the end of treatment.
Diary interaction compliance will be summarized with descriptive statistics and presented by treatment group.

Plasma concentrations of hormones and bone turnover markers will be summarized with descriptive statistics for each treatment group at each visit. Additionally, a within-subject change will be calculated as the post-baseline measurement minus the baseline measurement and summarized in the same way. Pharmacodynamic data and efficacy data may be evaluated by a population PD or population PK/PD approach. All details of population analyses will be described in a separate analysis plan and a separate report will be written. When deemed necessary, data from this study may be combined with data from other studies.

7.4.3 Analysis of Exploratory Endpoints

CCI [redacted]

Pharmacokinetics may be evaluated by a population PK approach. All details of population analyses will be described in a separate analysis plan and a separate report will be written. When deemed necessary, data from this study may be combined with data from other studies.

Weight and waist circumference will be summarized with descriptive statistics for each treatment group at Baseline and at Weeks 4, 8, 12, and 15. Additionally, a within-subject change will be calculated as the post-baseline measurement minus the baseline measurement and summarized in the same way.

The mean frequency of moderate to severe 12h day time and 12h night time vasomotor symptoms as well as change from baseline will be summarized by study week and for the Last on Treatment Week. See Section 7.10.6 for study week and endpoint derivations. Additionally, the analysis of exploratory efficacy endpoints will be based on the FAS. An analysis of covariance model as described in Section 7.4.1.1 will be used.

7.5 Analysis of Safety

All analyses of safety will be presented by treatment group for SAF, unless specified otherwise. For analyses by visit, the visit windows from Section 7.10.6 will be applied. Except for adverse events and follow-up visit data, only data up to one day after the last dose date will be considered for the analyses.

7.5.1 Adverse Events

Summaries and listings of SAEs and Serious TEAEs include SAEs upgraded by the sponsor based on review of the Sponsor's list of Always Serious terms if any upgrade was done.

The coding dictionary for this study will be MedDRA. It will be used to summarize AEs by SOC and PT.

An overview table will include the following details:

- Number of TEAEs,
- Number and percentage of subjects with TEAEs,
- Number of drug related TEAEs,
- Number and percentage of subjects with causally drug related TEAEs,
- Number of serious TEAEs,
- Number and percentage of subjects with serious TEAEs,
- Number of serious drug related TEAEs,
- Number and percentage of subjects with serious drug related TEAEs,
- Number of TEAEs leading to permanent discontinuation of study drug,
- Number and percentage of subjects with TEAEs leading to permanent discontinuation of study drug,
- Number and percentage of subjects with TEAEs leading to study drug interruption, and
- Number of deaths.

The number and percentage of subjects with TEAEs, as classified by SOC and PT will be summarized for each treatment group. Summaries will be provided for:

- TEAEs,
- drug related TEAEs,
- serious TEAEs,
- drug related serious TEAEs,

- TEAEs leading to permanent discontinuation of study drug,
- drug related TEAEs leading to permanent discontinuation of study drug,
- TEAEs excluding serious adverse events that equal to or exceed a threshold of 5% in any treatment group, and
- common TEAEs that equal to or exceed a threshold of 5% in any treatment group.

The number and percentage of subjects with TEAEs, as classified by PT only, will be summarized for each treatment group.

The number of TEAEs and the number and percentage of subjects with TEAEs, as classified by SOC and PT will also be summarized by severity and by relationship to study drug. In the subject count, if a subject has multiple TEAEs with the same SOC or PT, but with differing severity or relationship, then the subject will be counted only once with the worst severity and highest degree of relationship, however, if any of the severity or relationship values are missing then the subject will be counted only once with missing severity or relationship. In the adverse event count, the adverse events will be presented in each category they were classified to. Drug related TEAEs will be presented in a similar way by severity only.

Adverse events of special interest oral/facial paresthesia, uterine bleeding and elevation in ALT and/or AST >3 x ULN. A summary table and detailed listing of adverse events of special interest will be provided. A burden of toxicity graph will be generated for the adverse events of special interest.

7.5.2 Clinical Laboratory Evaluation

The baseline visit is the last measurement taken prior to initial study drug administration.

Quantitative clinical laboratory variables, i.e. hematology, biochemistry, and urinalysis will be summarized using mean, standard deviation, minimum, median, maximum, 10%, 90%, and lower and upper quartiles for each treatment group at each visit. Additionally, a within-subject change will be calculated as the post-baseline measurement minus the baseline measurement and summarized in the same way. Each laboratory result will be classified as low (L), normal (N), or high (H) at each visit according to the laboratory supplied reference ranges.

Shift tables will be generated for each treatment group at each visit.

Frequency tabulations of qualitative clinical laboratory variables (urinalysis) will be presented for each treatment group at each visit (excluding Visits 2A, 3A, and 4A).

7.5.2.1 Liver Enzymes and Total Bilirubin

The liver safety assessments will be summarized by the following categories below based on the measurements from Alkaline Phosphatase (ALP), Alanine Transaminase (ALT), total bilirubin, Aspartate Transaminase (AST), and their combinations are defined. These parameters will be measured from a central lab. The subject's highest value during the treatment period will be used.

Parameter	Potentially Clinically Significant Criteria
ALT	>3 x ULN
	>5 x ULN
	>8 x ULN
	>10 x ULN
	>20 x ULN
AST	>3 x ULN
	>5 x ULN
	>8 x ULN
	>10 x ULN
	>20 x ULN
ALT or AST	>3 x ULN
	>5 x ULN
	>8 x ULN
	>10 x ULN
	>20 x ULN
ALP	>1.5 x ULN
Total bilirubin	>1.5 x ULN
	>2 x ULN
ALT or AST and total bilirubin (*)	ALT or AST >3 x ULN and total bilirubin >2 x ULN
ALT or AST, ALP, and total bilirubin (*)	ALT or AST >3 x ULN, ALP <2 x ULN, and total bilirubin >2 x ULN

(*) Combination of values measured within the same day or up to one day apart.

The denominator for each criterion will be the number of subjects who have at least one value during the treatment period. The number and percentage of subjects meeting the critiera during the treatment and follow-up periods will be summarized by treatment group.

A detailed listing will be provided for subjects with suspected drug-induced liver injury (DILI).

7.5.3 Vital Signs

The baseline visit is the last measurement taken prior to initial study drug administration.

Vital signs (systolic blood pressure, diastolic blood pressure, pulse rate, and temperature) will be summarized using mean, standard deviation, minimum, median, maximum, 10%, 90%, and lower and upper quartiles by treatment group and visit. Additionally, a within-subject

change will be calculated per visit as the post-baseline measurement minus the baseline measurement and summarized by treatment group and visit.

7.5.4 Electrocardiograms (ECGs)

The baseline visit is the last measurement taken prior to initial study drug administration.

ECG variables will be summarized using (n), mean, standard deviation, minimum, median, maximum, 10%, 90%, and lower and upper quartiles for each treatment group at each treatment visit, including changes from baseline.

Number and percent of subjects with normal, not clinically significant abnormal, and clinically significant abnormal results as assessed by investigator for the 12 lead ECG will be tabulated by treatment group at each treatment visit. Local readings and central readings will be summarized separately.

The QT intervals (QT, QTcB, and QTcF) will be summarized using frequency tables for each treatment visit (excluding Visits 2A, 3A, and 4A) and time point for values of clinical importance using the range criteria below.

QT Interval Criteria (msec)	Description
≤450	Normal
>450	Borderline
>480	Prolonged
>500	Clinically significant

The QT intervals will also be summarized by the frequencies of subjects with a change from baseline of clinical importance using the criteria identified below. These summaries will be provided for each treatment visit and time point.

Parameter	Change from Baseline Criteria
QT Interval (msec)	<0
	≥0
	>30
	>60

Number and percent of subjects whose 12 lead ECG reading changed from normal at baseline to abnormal will be tabulated by treatment group for each visit.

7.5.5 Pregnancies

A detailed listing of all pregnancies will be provided.

7.5.6 Other Safety-Related Observations

Endometrial health assessment results from transvaginal ultrasound and endometrial biopsy will be listed.

Mammogram results, plasma bone density marker concentration, and C-SSRS results will be listed.

7.6 Analysis of PK

PK/PD analyses will be provided in a separate report.

7.7 Subgroups of Interest

The primary efficacy endpoint summaries will be repeated by smoking status (current vs. former/never). The subgroup analyses for CYP1A2 inducers / inhibitors may be performed if data is available.

7.8 Other Analyses

Pharmacogenomic samples will be collected. Data will be analyzed and reported separately.

7.9 Interim Analysis (and Early Discontinuation of the Clinical Study)

An early timepoint analysis is planned 4 weeks after the last randomization. The unblinded results will be restricted to a very small group independent of the study team. No changes to the study or the statistical analysis plan will be made as a result of this analysis (which is intended to assist with internal decision making only). Full details will be documented in a separate charter to ensure the integrity of the trial blinding; all study team personnel, including all those involved in data management, clinical, medical or statistical review of the data, will remain blinded until database hard-lock (see Section 3.3).

7.10 Handling of Missing Data, Outliers, Visit Windows, and Other Information

7.10.1 Missing Data

If there are no diary entries for a particular calendar day, and the patient reported receiving at least one morning dose of study medication or at least one evening dose of study medication on that day, the frequency of vasomotor symptoms will be imputed to 0; otherwise, the frequency of vasomotor symptoms will be considered missing.

If no vasomotor symptoms were recorded for the day, and the patient reported receiving at least one morning dose of study medication and at least one evening dose of study medication on that day, the severity will be 0.

Missing primary endpoints will be handled by multiple imputation for continuous variables, and non-responder imputation for binary variables.

For study drug compliance, the capsules returned will be imputed to 0 at the visit if no bottle was returned for a visit.

Multiple Imputation Method:

Multiple imputation will be used to impute missing data and will be implemented using SAS® PROC MI. Missing data may be the result of missing Week 4/Week 12 frequency/severity of vasomotor symptoms measurements or the result of patients discontinuing treatment prior to Week 4/Week 12. Missing at random will be assumed.

The imputation model will include the subject demographics (age [<median, ≥median], race [white, non-white], baseline weight [<median, ≥median], and smoking status), and baseline and post-baseline frequency/severity of vasomotor symptoms. Missing data will be imputed 100 times to generate 100 complete data sets. The seed will be 98765. The fully conditional specification method will be used. The 100 complete data sets will then be analyzed using the same analysis method as the one used to analyze the primary endpoint. See Section 7.4.1.1. The results from the 100 fitted models will be combined using SAS® PROC MIANALYZE.

If the model fails to converge, exclusion of subject demographics will be considered.

Safety Data Imputation:

As a general principle, no imputation of missing data for other variables will be done. Exceptions are the start and stop dates of AEs and concomitant medication in rare cases. The imputed dates will be used to allocate the concomitant medication and AEs to a treatment period, in addition to determining whether an AE is/is not treatment emergent. Listings of the AEs and concomitant medications will present the actual partial dates; imputed dates will not be shown.

7.10.2 Pooled Center

Approximately 352 subjects will be randomized at approximately 48 sites. When "center" is included as a factor in a statistical model, sites that have less than 10 FAS subjects in total will be identified and then combined for statistical analysis purposes according to the following algorithm:

- Step 1: Divide the sites into two groups with Group 1 including all sites that have at least 10 FAS subjects and Group 2 including all remaining sites. Sort each group in ascending order by total sample size and site number.
- Step 2: Starting at the top of the Group 2 list (i.e., the first site with the smallest total sample size), combine the minimum number of sites required to achieve a "pooled center" that has at least 10 FAS subjects in total. Continue forming "pooled centers" in this manner until all Group 2 sites have been grouped or it is no longer possible to form a "pooled center" with at least 10 FAS subjects in total.
- Step 3: If there is a site (or several sites) left after step 2, combine the site(s) with the last "pooled center" that is created. For the situation where no previous "pooled center" is created, combine the site(s) with the first site on the sorted Group 1 list.
- "Pooled center" will be assigned names PoolSite01, PoolSite02, etc. For sites not needing to be pooled, the original site name will be used such as Site001. For all sites that have been combined into pooled centers, the assigned pooled center will be used instead of the original site identification in all statistical models that include "center" as a factor. However, the original site identification will be used in all summaries of subject disposition or discontinuation by site and in all data listings.

7.10.3 Electronic and Paper Diaries

Subjects are to receive an electronic diary in which to record daily vasomotor symptoms during the study along with all home study drug intake. Paper diaries are to be used as back up to the electronic diary. In the case where there is an exact duplicate based on date and time between the electronic diary and the paper diary, the result from the electronic diary will be used. In the case where the date and/or time differs between electronic and paper diaries, results from both diaries will be used.

7.10.4 Last Dose Date

If the last dose date is missing from the eCRF, the last known dose date reported in the electronic diary will be used.

7.10.5 Outliers

All values will be included in the analyses.

7.10.6 Visit Windows

The study protocol gives the overall study schedule for these visits expressed as the number of days relative to Visit 2. Study Day will be calculated as date of visit/assessment – first dose date +1. There will be no Day 0. Visit 2 is scheduled to be Day 1 if the patient takes the first dose on Visit 2, and the day immediately prior to Visit 2 is Day -1.

Subjects will not be excluded from analyses due to the subject's failure to comply with the visit schedule.

If baseline is missing, the screening assessment can be used.

7.10.6.1 Efficacy Visit Windows

For the weeks on treatment (weeks prior to the follow-up period), the last day is always the last day of exposure to treatment.

The study week determination for the vasomotor symptoms data is based on the following:

Analysis study week	Actual assessment day
Baseline	7 non-missing days prior to D1
Week 1	D1 to D7
Week 2	D8 to D14
Week 3	D15 to D21
Week 4	D22 to D28
Week 5	D29 to D35
Week 6	D36 to D42
Week 7	D43 to D49
Week 8	D50 to D56
Week 9	D57 to D63
Week 10	D64 to D70
Week 11	D71 to D77
Week 12	D78 to D84 or last day of exposure to treatment, if this happened after day 84
Last on Treatment Week	Last 7 days of treatment
Follow-up Week 1	Day 1 FU – Day 7 FU ^
Follow-up Week 2	Day 8 FU – Day 14 FU ^
Follow-up Week 3	Day 15 FU – Day 21 FU ^

^ Day X FU is defined as last day of exposure to treatment + X

The assessment of a week will be derived if any information for 3 or more days was reported (morning dose, evening dose, or vasomotor symptom).

All other efficacy assessments, including those from unscheduled visits and regardless of visit label, will be allocated to analysis visits based on the table below:

CRF visit	Target day	Actual assessment day	Analysis visit
Visit 1	D-35	D-35 to D-1	Screening
Visit 2	D1	D1	Baseline
Visit 3	D29	D22 to D36	Week 4
Visit 4	D57	D50 to D64	Week 8
Visit 5	D85	D78 to Day 92	Week 12
End of treatment	Last dosing day	Closest observation to the date	EOT
Visit 6	D106	Last assessment after date of completion or termination of treatment period	Week 15 Follow-Up

The value which assessment day is the closest to the defined target day within these windows is used. If two values are equally close, the latter is used in the analysis.

7.10.6.2 Safety Visit Windows

For the weeks on treatment (weeks prior to the follow-up period), the last day is always the day after the last day of exposure to treatment.

All safety assessments, including those from unscheduled visits and regardless of visit label, will be allocated to analysis visits based on the table below:

CRF visit	Target day	Actual assessment day	Analysis visit
Visit 1	D-35	D-35 to D-1	Screening
Visit 2	D1	D1	Baseline
Visit 2A	D15	After first dose to D22	Week 2
Visit 3	D29	D23 to D36	Week 4
Visit 3A	D43	D37 to D50	Week 6
Visit 4	D57	D51 to D64	Week 8
Visit 4A	D71	D65 to D78	Week 10
Visit 5	D85	D79 to Day 92	Week 12
End of treatment	Last dosing day + 1 day	Closest observation to the date	EOT
Visit 6	D106	Last assessment after date of completion or termination of treatment period	Week 15 Follow-Up

The value which assessment day is the closest to the defined target day within these windows is used. If two values are equally close, the latter is used in the analysis.

8 CHANGES FROM PROTOCOL

Protocol Version	Changes added in SAP	Comment/rationale for change
Amendment 2: 05 March 2018	Weight and waist circumference will be added as an Exploratory Efficacy Endpoint.	Weight and waist circumference will be analyzed as exploratory efficacy endpoints.
Amendment 2: 05 March 2018	The day time vasomotor symptoms and night time vasomotor symptoms of the 4 co-primary endpoints will be added as an Exploratory Efficacy Endpoint.	Day time vasomotor symptoms and night time vasomotor symptoms of the 4 co-primary endpoints will be analyzed as exploratory efficacy endpoints.

9 DOCUMENT REVISION HISTORY

Version	Date	Changes	Comment/rationale for change
1.0	DD-Mmm-YYYY	N/A	Document finalized

10 REFERENCES

1. US Department of Health and Human Services FDA. Guidance for Industry estrogen and estrogen/progestin drug products to treat vasomotor symptoms and vulvar and vaginal atrophy symptoms – Recommendations for Clinical Evaluation. (2003).
2. European Medicines Agency. Guideline of clinical investigation of medicinal products for hormone replacement therapy of oestrogen deficiency symptoms in menopausal women (2005).
3. National Institutes of Health. Assessing and improving measures of hot flashes. (2004).
4. Burgess, *et al.* Baseline data in clininical trials. *MJA* Vol 179, 21 Jul 2003.
5. Carpenter, J. S. The Hot Flash Related Daily Interference Scale: a tool for assessing the impact of hot flashes on quality of life following breast cancer. *J. Pain Symptom Manage.* **22,** 979-89 (2001).
6. Daut, R. L., Cleeland, C. S. & Flanery, R. C. Development of the Wisconsin Brief Pain Questionnaire to assess pain in cancer and other diseases. *Pain* **17,** 197–210 (1983).
7. Hann, D. M. *et al.* Measurement of fatigue in cancer patients: development and validation of the Fatigue Symptom Inventory. *Qual. Life Res.* **7,** 301–10 (1998).
8. Rand, K. L, et al. Modeling hot flushes and quality of life in breast cancer survivors. *Climacteric* **14,** 171-80 (2011).
9. Parrott, A. C. & Hindmarch, I. Factor analysis of a sleep evaluation questionnaire. *Psychol. Med.* **8,** 325-329 (1978).
10. Greene, J. G. A factor analytic study of climacteric symptoms. *J. Psychosom. Res.* **20,** 425-430 (1976).
11. Hilditch, J., Lewis, J. & Peter, A. A menopause-specific quality of life questionnaire: development and psychometric properties. *Maturitas* **24,** 161-175 (1996).
12. Lewis, J. E., Hilditch, J. R. & Wong, C. J. Further psychometric property development of the Menopause-Specific Quality of Life questionnaire and development of a modified version, MENQOL-Intervention questionnaire. *Maturitas* **50,** 209-221 (2205).
13. Radtke, J., Terhorst, L. & Cohen, S. The Menopause-Specific Quality of Life (MENQOL) Questionnaire: Psychometric evaluation among breast cancer survivors. *Menopause* **18,** 289-295 (2011).

11 APPENDICES

11.1 Appendix 1: Key Contributors and Approvers

List of Key Contributors and Approvers

Key Contributors

The following contributed to or reviewed this Statistical Analysis Plan as relevant to their indicated discipline or role.

Primary author (s)

PPD	Medpace Biostatistics

Contributors and Reviewers

PPD	Medpace Biostatistics
PPD	Medpace Medical Affairs
PPD	Astellas Biostatistics
PPD	Astellas Biostatistics
PPD	Ogeda, S.A.
PPD	Astellas Pharma Europe B.V.
PPD	Astellas Pharma Europe B.V
PPD	Astellas Pharma Europe B.V
PPD	Astellas Medical and Development

Author and Approver Signatories

This Statistical Analysis Plan was approved by: