Study Title:	LIBERTY 1: An International Phase 3 Randomized, Double- Blind, Placebo-Controlled Efficacy and Safety Study to Evaluate Relugolix Co-Administered with and without Low-Dose Estradiol and Norethindrone Acetate in Women with Heavy Menstrual Bleeding Associated with Uterine Fibroids LIBERTY 2: An International Phase 3 Randomized, Double-Blind, Placebo-Controlled Efficacy and Safety Study to Evaluate Relugolix Co-Administered with and without Low-Dose Estradiol and Norethindrone Acetate in Women with Heavy Menstrual Bleeding Associated with Uterine Fibroids
NCT Number:	03103087
Document Date:	Statistical Analysis Plan Amendment 1: 14-Jun-2019

16.1.9. Documentation of Statistical Methods

STATISTICAL ANALYSIS PLAN

Study Titles:	LIBERTY 1: An International Phase 3 Randomized, Double-Blind, Placebo-Controlled Efficacy and Safety Study to Evaluate Relugolix Co-Administered with and without Low-Dose Estradiol and Norethindrone Acetate in Women with Heavy Menstrual Bleeding Associated with Uterine Fibroids LIBERTY 2: An International Phase 3 Randomized, Double-Blind, Placebo-Controlled Efficacy and Safety Study to Evaluate Relugolix Co-Administered with and without Low-Dose Estradiol and Norethindrone Acetate in Women with Heavy Menstrual Bleeding Associated with Uterine Fibroids
Investigational Product:	Relugolix
Protocol Number:	MVT-601-3001 and MVT-601-3002
Indication:	Heavy menstrual bleeding associated with uterine fibroids
Sponsor:	Myovant Sciences GmbH Viaduktstrasse 8 4051 Basel Switzerland
Regulatory Identifier(s):	IND # 131161 EudraCT # 2016-003727-27
Version/Effective Date:	Original: 07-May-2019 Amendment 1: 14-Jun-2019

STATISTICAL ANALYSIS PLAN APPROVAL SHEET

MVT-601-3001 (LIBERTY 1): An International Phase 3 Randomized, Double-Blind, Placebo-Controlled Efficacy and Safety Study to Evaluate Relugolix Co-Administered with and without Low-Dose Estradiol and Norethindrone Acetate in Women with Heavy Menstrual Bleeding Associated with Uterine Fibroids

MVT-601-3002 (LIBERTY 2): An International Phase 3 Randomized, Double-Blind, Placebo-Controlled Efficacy and Safety Study to Evaluate Relugolix Co-Administered with and without Low-Dose Estradiol and Norethindrone Acetate in Women with Heavy Menstrual Bleeding Associated with Uterine Fibroids

This statistical analysis plan has been approved by Myovant Sciences GmbH ("Myovant"), with Myovant Sciences, Inc., acting as agent of Myovant. The following signatures document this approval.

PPD

14 Jun 2019
Date

14 Jun 2019
Date

14 Jun 2019
Date

14 JUN 2019
Date

14 Jun 2019
Date

PPD

14 JUN - 2019
Date

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Term	Definition/Explanation
ALP	alkaline phosphatase
ALT	alanine aminotransferase
ANOVA	analysis of variance
AST	aspartate aminotransferase
ATC	Anatomical Therapeutic Chemical
AUC	area under the curve
BMD	bone mineral density
BMI	body mass index
C_τ	predose trough concentrations
CDF	cumulative distribution function
CFI	comparative fit index
CI	confidence interval
CRF	case report form
CSR	clinical study report
CTCAE	common terminology criteria for adverse events
DSMB	data safety monitoring board
DXA	dual-energy x-ray absorptiometry
E2	estradiol
ECG	electrocardiogram
eCRF	electronic case report form
EDC	electronic data capture
eDiary	electronic diary
EOT	end-of-treatment
EQ-5D-5L	European Quality of Life Five-Domain Five-Level
FP	feminine product
FPRR	feminine product return rate
FSH	follicle-stimulating hormone
GFI	goodness of fit index
Hgb	hemoglobin
ICH	International Council on Harmonisation
ITT	intent-to-treat
KM	Kaplan Meier
LH	luteinizing hormone
LLN	lower limit of normal
LS	least squares
max	maximum
MBL	menstrual blood loss
min	minimum
mITT	modified intent to treat

Term	Definition/Explanation
MedDRA	Medical Dictionary for Regulatory Activities
mmHg	millimeters of mercury
M-vol	myoma volume
NET	norethindrone
NETA	norethindrone acetate
NRS	Numerical Rating Scale
PBO	placebo
PDF	probability density function
PGA	patient global assessment
PK	pharmacokinetic
PT	Preferred Term
QD	once daily
QTcF	corrected QT interval Fridericia
RMSEA	root mean square error of approximation
SAP	statistical analysis plan
SD	standard deviation
SES	standardized effect size
SMQ	standard MedDRA query
SOC	System Organ Class
UFS-QoL	Uterine Fibroid Symptom and Health-Related Quality of Life (Questionnaire)
ULN	upper limit of normal
U-vol	uterine volume
WHO	World Health Organization
Wks	weeks

1. INTRODUCTION

The purpose of this statistical analysis plan (SAP) is to describe the analyses planned for phase 3 studies MVT-601-3001 (LIBERTY 1) and MVT-601-3002 (LIBERTY 2), both entitled "An International Phase 3 Randomized, Double-Blind, Placebo-Controlled Efficacy and Safety Study to Evaluate Relugolix Co-Administered with and without Low Dose Estradiol and Norethindrone Acetate in Women with Heavy Menstrual Bleeding Associated with Uterine Fibroids." In these studies, patients are randomized to one of three treatment arms: relugolix 40 mg + estradiol/norethindrone acetate (E2/NETA) 1 mg/0.5 mg for 24 weeks (Group A, also referred to as the relugolix + E2/NETA group), relugolix 40 mg for 12 weeks followed by 12 weeks of relugolix 40 mg + E2/NETA 1 mg/0.5 mg (Group B, also referred to as the relugolix + delayed E2/NETA group), or placebo for 24 weeks (Group C, also referred to as the placebo group).

The 2 phase 3 studies are replicative; the only difference between the two protocols is the Week 24 endometrial biopsies, which in MVT-601-3001 are done in all patients and in MVT-601-3002 depend on the results of the Week 24 ultrasound.

This SAP was developed in accordance with the International Council on Harmonisation (ICH) E9 guidelines. All decisions regarding statistical analysis of the study, as defined in this SAP, will be made prior to unblinding of the study data.

The SAP is based on:

- Protocol MVT-601-3001, Amendment 2, dated 18 Sept 2017;
- Protocol MVT-601-3002, Amendment 2, dated 25 Sept 2017;
- ICH guidelines E3 (Clinical Study Reports) and E9 (Statistical Principles for Clinical Trials).

This document may evolve over time (eg, to reflect the requirements of protocol amendments or regulatory requests). However, the SAP is to be finalized, approved by the sponsor, and placed on file before the database is locked. Changes to the final approved plan will be noted in the clinical study report (CSR). Unless otherwise specified, the objectives, definitions of endpoints, and pre-specification of analyses presented in this document apply to both studies.

1.1. Study Objectives and Endpoints

The study objectives and corresponding endpoints are listed in the following table. The endpoints in *italics* are not listed in the protocol, but they have been identified as important for assessment of treatment effect on the basis of emerging data and clinical relevance to the study objectives and therefore are included in this SAP.

Table 1: Study Objectives and Endpoints

Objective(s)	Endpoint(s)
Primary Efficacy	
To determine the benefit of relugolix 40 mg once daily co-administered with E2 1 mg and NETA 0.5 mg compared with placebo for 24 weeks on heavy menstrual bleeding associated with uterine fibroids	Proportion of women in the relugolix + E2/NETA group versus the placebo group who achieve an MBL volume of < 80 mL AND at least a 50% reduction from Baseline MBL volume over the last 35 days of treatment, as measured by the alkaline hematin method
Key Secondary Efficacy (Alpha-Protected for Hierarchical Hypothesis Testing — relugolix + E2/NETA versus placebo)	
Achievement of amenorrhea	Proportion of women who achieve amenorrhea over the last 35 days of treatment
Heavy menstrual bleeding associated with uterine fibroids	Percent change from Baseline to Week 24 in MBL volume
Impact of uterine fibroids on symptoms, activities, and health-related quality of life as measured by components of the UFS-QoL	*Change from Baseline to Week 24 in the UFS-QoL Bleeding and Pelvic Discomfort Scale score, a sub-scale of the UFS-QoL Symptom Severity scale*
Change in hemoglobin	*Proportion of women with a hemoglobin ≤ 10.5 g/dL at Baseline who achieve an increase of > 2 g/dL from Baseline to Week 24*
Pain associated with uterine fibroids	*Proportion of patients with a maximum NRS score ≤ 1 during the last 35 days before the last dose of study drug in the subset of women with a maximum NRS score ≥ 4 for pain associated with uterine fibroids during the last 35 days prior to randomization*
Uterine fibroid volume	Percent change from Baseline to Week 24 in uterine fibroid volume
Uterine volume	Percent change from Baseline to Week 24 in uterine volume
Other Secondary Efficacy (Not for Hierarchical Hypothesis Testing) [a]	
To determine the benefit of relugolix 40 mg once daily for 12 weeks followed by 12 weeks of relugolix 40 mg once daily co-administered with E2 1 mg and NETA 0.5 mg compared with placebo for 24 weeks on heavy menstrual bleeding associated with uterine fibroids	Proportion of women in the relugolix + delayed E2/NETA group versus the placebo group who achieve an MBL volume of < 80 mL AND at least a 50% reduction from Baseline MBL volume over the last 35 days of treatment, as measured by the alkaline hematin method
Heavy menstrual bleeding associated with uterine fibroids	• *Percent change from Baseline in MBL volume by visit* • *Change from Baseline in MBL volume by visit*

Objective(s)	Endpoint(s)
	• Time to achieve an MBL volume of < 80 mL AND at least a 50% reduction from Baseline MBL volume as measured by the alkaline hematin method • *Proportion of women in the relugolix + E2/NETA group versus the placebo group who achieve an MBL volume of < 80 mL AND at least a 50% reduction from Baseline MBL volume by visit*
Achievement of amenorrhea	• *Sustained amenorrhea rate by visit* • *Time to achieving sustained amenorrhea* • *Time to achieving amenorrhea*
Change in hemoglobin	• Proportion of women with a hemoglobin below the lower limit of normal at Baseline who achieve an increase of ≥ 1 g/dL from Baseline at Week 24 • *Change from Baseline to Week 24 in hemoglobin for women with a hemoglobin ≤ 10.5g/dL at Baseline*
Impact of uterine fibroids on symptoms, activities and health-related quality of life as measured by components of the UFS-QoL	• Change from Baseline to Week 24 in the UFS-QoL Symptom Severity Scale score • Change from Baseline to Week 24 in the UFS-QoL Activities Scale score • *Change from Baseline to Week 24 in the UFS-QoL Revised Activities Scale score* • *Proportion of responders who achieved a meaningful increase of at least 20 points from Baseline to Week 24 in UFS-QoL Revised Activities Scale score* • *Proportion of responders who achieved a meaningful reduction of at least 20 points from Baseline to Week 24 in UFS-QoL Bleeding and Pelvic Discomfort Scale score* • Change from Baseline to Week 24 in the interference of uterine fibroids with physical activities based on UFS-QoL Question 11 • Change from Baseline to Week 24 in the interference of uterine fibroids with social activities based on UFS-QoL Question 20

Objective(s)	Endpoint(s)
	• Change from Baseline to Week 24 in embarrassment caused by uterine fibroids based on UFS-QoL Question 29
Patient global assessment for function and symptoms as measured by the PGA for function and symptoms	• Change in PGA for uterine fibroid related function from Baseline to Week 24 • Change in PGA for uterine fibroid symptoms from Baseline to Week 24 • *Proportion of patients achieving improvement from Baseline in PGA for uterine fibroid symptoms from Baseline to Week 24* • *Proportion of patients achieving improvement from Baseline in PGA for uterine fibroid related function from Baseline to Week 24*
Impact of heavy menstrual bleeding on social, leisure, and physical activities as measured by the Menorrhagia Impact Questionnaire	• Change from Baseline to Week 24 in the Menorrhagia Impact Questionnaire Score for physical activities • Change from Baseline to Week 24 in the Menorrhagia Impact Questionnaire Score for social and leisure activities
Pain associated with uterine fibroids[b]	Proportion of women who achieve a *maximum* NRS score for pain associated with uterine fibroids over the last 35 days of treatment that is at least a 30% reduction from Baseline in the subset of women with a maximum pain score ≥ 4 during the 35 days prior to randomization
Safety	
To determine the safety of 24 weeks of relugolix 40 mg once daily co-administered with either 12 or 24 weeks of E2 1 mg and NETA 0.5 mg in women with heavy menstrual bleeding associated with uterine fibroids compared with placebo for 24 weeks	Treatment-emergent adverse events, change in vital signs (including weight), clinical laboratory tests, and electrocardiograms
To determine the percent change from Baseline to Week 12 in bone mineral density at the lumbar spine (L1-L4) in the relugolix + E2/NETA group compared with the relugolix + delayed E2/NETA group in women with heavy menstrual bleeding associated with uterine fibroids	Percent change from Baseline to Week 12 in bone mineral density at the lumbar spine (L1-L4) in the relugolix + E2/NETA group compared with relugolix + delayed E2/NETA group as assessed by DXA

Objective(s)	Endpoint(s)
To determine the change in bone mineral density of women with heavy menstrual bleeding associated with uterine fibroids treated with 24 weeks of relugolix 40 mg once daily co-administered with either 12 or 24 weeks of E2 1 mg and NETA 0.5 mg compared with placebo for 24 weeks	Percent change from Baseline to Week 24 in bone mineral density at the lumbar spine (L1-L4), total hip, and femoral neck as assessed by DXA
To determine the incidence of vasomotor symptoms with relugolix 40 mg once daily co-administered with either 12 or 24 weeks of E2 1 mg and NETA 0.5 mg in women with heavy menstrual bleeding associated with uterine fibroids	Incidence of vasomotor symptoms
Pharmacokinetic and Pharmacodynamic	
To evaluate the pharmacokinetic and pharmacodynamic effects of 24 weeks of relugolix 40 mg once daily when co-administered with either 12 or 24 weeks of E2 1 mg and NETA 0.5 mg	• Predose trough concentrations (C_τ) of relugolix, and NET and Baseline-adjusted E2 concentration • Absolute and changes from Baseline to Week 24 in predose concentrations of LH, FSH, E2, and progesterone
Exploratory	
To determine the benefit of 24 weeks of relugolix 40 mg once daily co-administered with either 12 or 24 weeks of E2 1 mg and NETA 0.5 mg compared with placebo on patient-reported quality of life outcome measures (EQ-5D-5L)	Change from Baseline to Week 24 in the EQ-5D-5L Scale score

Abbreviations: DXA, dual energy x-ray absorptiometry; E2, estradiol; EQ-5D-5L, European Quality of Life Five-Domain Five-Level; FSH, follicle-stimulating hormone; LH, luteinizing hormone; MBL, menstrual blood loss; NET, norethindrone; NETA, norethindrone acetate; NRS, numerical rating scale; PGA, Patient Global Assessment; UFS-QoL, Uterine Fibroid Symptom and Health-Related Quality of Life.

[a] The secondary endpoints below will be assessed comparing the relugolix + E2/NETA group with the placebo group inferentially; the relugolix + E2/NETA group to the relugolix + delayed E2/NETA group and the relugolix + delayed E2/NETA group to the placebo group descriptively, unless otherwise specified.

[b] Changed from mean NRS score (in the protocol) to maximum NRS score. Since pain associated with uterine fibroids is mostly during menstrual days, mean NRS scores over the last 35 days is very low (< 1) for most patients, hence, not appropriate to define percent reduction from Baseline.

2. STUDY DESIGN

2.1. Summary of Study Design

The LIBERTY 1 and LIBERTY 2 studies are two replicate, randomized, double-blind, placebo-controlled phase 3 studies evaluating the efficacy and safety of relugolix 40 mg in combination with E2 1 mg/NETA 0.5 mg in women with heavy menstrual bleeding associated with uterine fibroids (MVT-601-3001, MVT-601-3002). Patients with heavy menstrual bleeding associated with uterine fibroids — as evidenced by a menstrual blood loss (MBL) volume of ≥ 80 mL per cycle for 2 cycles or ≥ 160 mL during one cycle, as measured by the alkaline hematin method during the screening period — who met other eligibility criteria were randomly assigned (1:1:1) to 1 of the 3 treatment arms:

- Group A (relugolix + E2/NETA): relugolix 40 mg once daily co-administered with E2 1 mg/NETA 0.5 mg for 24 weeks;
- Group B (relugolix + delayed E2/NETA): relugolix 40 mg once daily for 12 weeks followed by relugolix 40 mg once daily co-administered with E2 1 mg/NETA 0.5 mg for 12 weeks;
- Group C (placebo): placebo for 24 weeks

Randomization was stratified as follows:

- Geographic Region: North America versus Rest of World;
- Mean screening MBL volume using alkaline hematin method: < 225 mL versus ≥ 225 mL.

The primary endpoint for both trials is the proportion of women receiving relugolix + E2/NETA (Group A) versus placebo (Group C) who achieve BOTH a MBL volume of < 80 mL AND at least a 50% reduction from Baseline in MBL volume over the last 35 days of treatment, as measured by the alkaline hematin method.

This study includes a screening period (up to ~13 weeks), a randomized treatment period (24 weeks), and a safety follow-up period (~30 days). During the screening period, diagnoses of uterine fibroids are confirmed by centrally reviewed transvaginal ultrasound. Women with iron-deficient microcytic anemia and hemoglobin ≥ 8 g/dL and ≤ 10 g/dL during the screening period are treated with oral or parenteral iron replacement therapy. After randomization, patients begin double-blinded study drug treatment for 24 weeks.

Patients who complete LIBERTY 1 or LIBERTY 2, including those randomized to placebo, and who meet other eligibility criteria are offered the opportunity to enroll in a 28-week open-label extension study, in which all patients will receive relugolix 40 mg co-administered with E2 1 mg and NETA 0.5 mg. Patients who do not enroll into the extension study have a safety follow-up visit approximately 30 days after their last doses of study medication.

Additional safety follow-up may be performed after the safety follow-up visit. Data collected during the additional safety follow-up period will be summarized and reported in an addendum to the respective clinical study report. Patients who are not proceeding into the extension study and who have endometrial hyperplasia or endometrial cancer on the endometrial biopsy should be treated as per standard of care and additional follow-up should be evaluated and managed, as

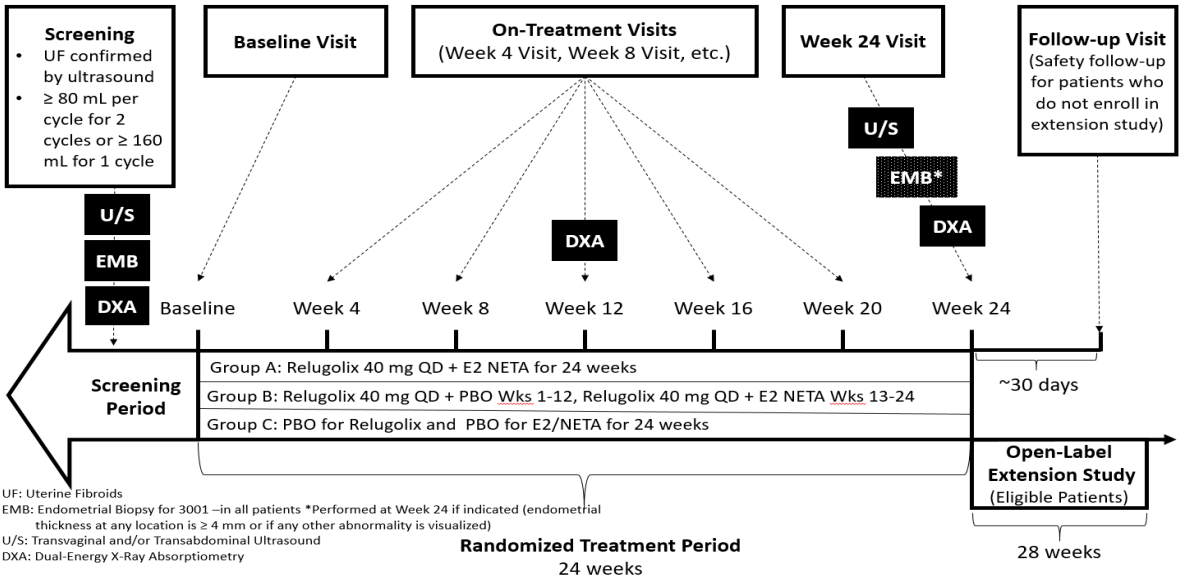
needed, by a gynecologist. In addition, they should undergo a repeat biopsy in 3 to 6 months after the Week 24/Early Termination and will be contacted to obtain information on procedures performed or treatments received (if any) for the biopsy findings through the time of the repeat biopsy. The repeat biopsy will be submitted to the central laboratory.

Patients who are not proceeding into the extension study and who have a bone mineral density (BMD) loss of > 2% at the lumbar spine (L1–L4) or total hip relative to the Baseline measurement at their Week 24/Early Termination visit will undergo a follow-up DXA scan 6 months (± 1 month) after discontinuation of study drug and will be contacted to obtain information about medications and conditions (eg, pregnancy, hyperparathyroidism, hypothyroidism, etc.) that might affect BMD through the time of the repeat DXA scan. If the DXA scan 6 months post-treatment continues to show BMD loss of > 1.5% at the lumbar spine and/or > 2.5% at the total hip compared with Baseline, patients will have an additional scan at 12 months post-treatment. All follow-up DXA scans will be submitted for central reading. Patients whose menses had not resumed as of the safety follow-up visit for unexplained reasons will be contacted by telephone to determine if menses have resumed. Patients with reductions in visual acuity will be referred for ophthalmology consultation.

An external independent data and safety monitoring board (DSMB) was established to review periodic safety analyses, including BMD assessments. The roles and responsibilities of the independent DSMB are described in a separate charter. A separate SAP was created to document the specific safety data analyses that would be performed by an independent data coordinating center for the DSMB on an ongoing basis during the study.

A schematic of the study is presented in Figure 1.

Figure 1: Study Schematic



Abbreviations: E2, estradiol; NETA, norethindrone; PBO, placebo; QD, once daily; Wks, weeks.

2.2. Sample Size Considerations

2.2.1. Sample Size Justifications for Primary Efficacy Endpoint

The following assumptions were used to determine the sample size for this study:

- 2-sided type I error rate: 0.05
- Randomization: 1:1:1
- Responder rate for placebo group: 25%
- Difference in responder rates between the relugolix + E2/NETA group and the placebo group: 30%
- Dropout rate: ~20%

With the assumption of a dropout rate of 20%, approximately 130 women in the relugolix + E2/NETA group and 130 women in the placebo group will provide at least 99% power at a 2-sided 0.05 significance level to detect a 30% difference in responder rates between relugolix + E2/NETA group and the placebo group for the primary endpoint. With an additional 130 women in the relugolix + delayed E2/NETA group, the total sample size will be approximately 390 women.

The assumed responder rate of 25% for the placebo group is within the range of responder rates observed from similar phase 3 trials in uterine fibroids (Stewart, 2017). The sample size and power calculations are based on a chi-squared test.

2.2.2. Sample Size Justifications for Percent Change in Bone Mineral Density at 12 Weeks

A pooled analysis of the percent change in BMD at 12 weeks using data from both phase 3 studies is described separately in the statistical analysis plan for the Integrated Summary of Safety. The results of this pooled analysis comparing the relugolix + E2/NETA group with the relugolix + delayed E2/NETA group will be presented in the Integrated Summary of Safety and will not be included in the CSRs for these studies.

For the comparison of the relugolix + E2/NETA group with the relugolix + delayed E2/NETA group with respect to the percent change in BMD from Baseline to Week 12 at the lumbar spine (L1–L4), approximately 260 women in the relugolix + E2/NETA group (pooled between the LIBERTY 1 and LIBERTY 2 studies) and 260 women in the relugolix + delayed E2/NETA (pooled) will provide at least 90% power at a 2-sided 0.05 significance level to detect a 1.25% absolute treatment difference, assuming a standard deviation of 4% and up to 15% dropout rate for each treatment group. Power calculations for this BMD comparison are based on a two-sample t-test.

Sample size and power calculations were performed using the software package *nQuery* 4.0 (Statistical Solutions Ltd.).

3. PLANNED ANALYSES

3.1. Interim Analyses

No interim efficacy analyses were planned or performed for these two studies.

An external, independent DSMB was established to review periodic safety analyses, including BMD assessments. A separate SAP was created to document the specific safety data analyses that would be performed by an independent data coordinating center for the DSMB on an ongoing basis during the study.

3.2. Final Analyses

The final analysis of all efficacy and safety data from MVT-601-3001 and MVT-601-3002 will occur after approximately 390 patients have been randomized to each study and have had the opportunity to be followed for 24 weeks of study treatment and through the 30-day safety follow-up visit. This document describes this final analysis.

There will be periodic safety data review by the DSMB. An independent data coordinating center has performed the periodic safety analyses and has provided results of these analyses to the DSMB, as defined in the DSMB charter and outlined in a separate DSMB SAP.

3.3. Safety Follow-Up Analyses

Patients who are not proceeding into the extension study and who have endometrial hyperplasia or endometrial cancer on the endometrial biopsy should be treated as per standard of care and additional follow up should be evaluated and managed, as needed, by a gynecologist. In addition, they should undergo a repeat biopsy in 3 to 6 months after the Week 24/Early Termination and will be contacted to obtain information on procedures performed or treatments received (if any) for the biopsy findings through the time of the repeat biopsy. The repeat biopsy will be submitted to the central laboratory.

Patients who are not proceeding into the extension study and who have a BMD loss of > 2% at the lumbar spine (L1–L4) or total hip relative to the Baseline measurement at their Week 24/Early Termination visit will undergo a follow-up DXA scan 6 months (± 1 month) after discontinuation of study drug and will be contacted to obtain information about medications and conditions (eg, pregnancy, hyperparathyroidism, hypothyroidism, etc) that might affect bone mineral density through the time of the repeat DXA scan. If the DXA scan 6 months post-treatment continues to show BMD loss of > 1.5% at the lumbar spine and/or > 2.5% at the total hip compared to Baseline patients will have an additional scan at 12 months post-treatment. All follow-up DXA scans will be submitted for central reading. Patients whose menses had not resumed as of the safety follow-up visit for unexplained reasons will be contacted by telephone to determine if menses have resumed. Patients with reductions in visual acuity will be referred for ophthalmology consultation.

Data collected during the additional safety follow-up period will be summarized and reported in an addendum to the respective clinical study report.

4. GENERAL CONSIDERATIONS FOR DATA ANALYSES AND HANDLING OF MISSING DATA

4.1. Data Presentation Conventions

All statistical analyses will be conducted using SAS® Version 9.2 or higher.

A statistical test for the primary and secondary efficacy endpoints will be assessed at a two-sided $\alpha = 0.05$ significance level, and all confidence intervals (CIs) will be reported as two-sided unless otherwise stated.

Where appropriate, variables will be summarized descriptively by study visit. For the categorical variables, the count and proportions of each possible value will be tabulated by treatment group. For continuous variables, the number of patients with non-missing values, mean, median, standard deviation (SD), minimum, and maximum values will be tabulated.

Unless otherwise specified, the following conventions will be applied to all analyses:

- Mean and median values will be formatted to one more decimal place than the measured value. Standard deviation values will be formatted to two more decimal places than the measured value; minimum and maximum values will be presented to the same number of decimal places as the measured value; if the measured value is large (eg, > 100), fewer decimal places may be displayed.
- Percentages will be rounded to 1 decimal place;
- p-values will be rounded to 4 decimal places. p-values < 0.0001 will be presented as "< 0.0001" and p-values > 0.9999 will be presented as "> 0.9999";
- 1 month = 30.4375 days. Month is calculated as (days/30.4375) rounded to 1 decimal place;
- 1 year = 365.25 days. Year is calculated as (days/365.25) rounded to 1 decimal place;
- Age will be calculated using the date of randomization. If only year of birth is collected, 1 July of the year of birth will be used to calculate age.
- 1 pound = 0.454 kg;
- 1 inch = 2.54 cm;
- Missing efficacy or safety data will not be imputed unless otherwise specified;
- For laboratory results above or below sensitivity limits displayed as "<" or ">" a quantification threshold, 0.0000000001 will be subtracted or added, respectively, to the threshold to derive a numeric result for analyses;
- For MBL volume reported as below the limit of quantification (for example, MBL below Quantification Level <5.0 mL or <2.5 mL), 0.0000000001 will be subtracted from the reported quantification threshold for the visit to derive a numeric result for analyses;
- For safety analyses, calculation of percentages will be calculated on the basis of the number of patients in the analysis population in each treatment group;

- For by-visit observed data analyses, calculation of percentages will be calculated on the basis of the number of patients with non-missing data as the denominator, unless otherwise specified;
- For other continuous endpoints, the summary statistics will include mean, SD, median, and range (minimum and maximum);
- For time-to-event endpoints, the summary statistics will include median time to event-free survival, 25th and 75th percentiles and number of patients at risk at specified time points;
- For categorical endpoints, the summary statistics will include counts and percentages;
- Confidence intervals, when presented, will generally be constructed at the 95% level. For binomial variables, exact methods will be employed, unless otherwise specified.

4.2. Analysis Populations

Three analysis populations are defined below. Number and percent of patients meeting the definition of each analysis population will be summarized by treatment group.

4.2.1. Modified Intent-to-Treat Population

Efficacy analyses will be performed using the modified Intent-to-Treat (mITT) population, unless otherwise specified. The mITT population is defined as all randomized patients who have received any amount of study drug (relugolix/placebo or E2/NETA/placebo). Efficacy analyses will be performed by treatment group as randomized.

4.2.2. Per-Protocol Population

The Per-Protocol population will consist of those members of the mITT population who do not have any of the specified subset of important protocol deviations (see Section 5.3).

The Per-Protocol population will not be analyzed if this population comprises > 95% or < 50% of the mITT population. The Per-Protocol population will be used for sensitivity analysis of the primary efficacy endpoint. The Per-Protocol population and the associated subset of important protocol deviations will be identified prior to unblinding the trial.

4.2.3. Safety Population

Safety analyses will be performed using the Safety population unless otherwise specified. The Safety population is the same as the mITT population and is defined as all randomized patients who have received any amount of study drug. Safety data will be analyzed by treatment group according to the actual treatment received (not the randomized treatment). Any patient who received at least one dose of relugolix will be considered as a relugolix patient.

4.3. Definitions, Computation, and Convention

4.3.1. Definition of Date of First Dose and Date of Last Dose of Study Drug

The date of the first dose of study drug is defined as the date when a patient receives the first dose of study drug (relugolix/placebo or E2/NETA/placebo). The date of the last dose of study

drug is defined as the date a patient receives the last dose of study drug. If the complete date of last dose of study drug is unknown, the last date the study drug was known to have been taken will be used.

4.3.2. Study Day

Study day will be calculated with respect to the date of the first dose of study drug (Study Day 1). For assessments conducted on or after the date of the first dose of study drug, study day will be calculated as:

(Assessment date – date of first dose of study drug) + 1

For assessments conducted before the date (and time) of the first dose of study drug, study day will be calculated as:

(Assessment date – date of first dose of study drug)

For patients who do not receive any amount of study drug, study day will be calculated as above with respect to the date of randomization.

4.3.3. Definition of Treatment Duration

Treatment duration is defined as the duration of time from the date of the first dose of study drug to the date of the last dose of study drug as follows:

(Date of last dose of study drug – Date of first dose of study drug) + 1

For patients without complete date of last dose of study drug, the last date study drug was known to have been taken will be used to calculate treatment duration. For patients who did not return for the Early Termination visits, the time after their last visit will not be included in calculations of treatment duration.

4.3.4. Definition of Baseline Value and Post-Baseline Value

Unless otherwise specified, Baseline values are defined as the last measurement before the first administration (date and time) of study drug. A post-Baseline value is defined as a measurement taken after the first administration of study drug. Change from Baseline is defined as (post-Baseline value – Baseline value). Both date and time of study drug administration and measurement will be considered when calculating Baseline value. If the time is not available, then the date alone will be used. For patients who receive no study medication, the date of randomization will be used in place of the date of first dose in determining Baseline and post-Baseline values.

4.3.5. Visit Windows

Visit windows, which will be used to associate assessments with a scheduled visit, will be used only for summarizing data by visit. The windows for scheduled assessments are shown in Table 2, Table 3 (electrocardiogram [ECG], BMD, Uterine Fibroid Symptom and Health-Related Quality of Life [UFS-QoL]), and Table 4 (transvaginal ultrasound, endometrial biopsy, and European Quality of Life Five-Domain Five-Level [EQ-5D-5L]), respectively. For both efficacy and safety assessments, the study day will be used to determine the associated visit window.

The data collected in the electronic diary (eDiary) related to bleeding and use of feminine products will be assigned to visit windows as specified in Table 5 and will be used to calculate the feminine product return rate (FPRR) as specified in Section 7.3.3.

If the results from more than one monthly or Week 12/Week 24 assessment are within a given visit window, the non-missing result from the assessment closest to the target date will be used. If two assessments are equally close to the target day, the earlier assessment will be used. For summaries of shift from Baseline in safety parameters, all values will be considered for these analyses.

Table 2: Visit Windows for Monthly Assessments

Visit	Start Day	Target Day	End Day
Week 4[a]	1	29	43
Week 8	44	57	71
Week 12	72	85	99
Week 16	100	113	127
Week 20	128	141	155
Week 24	156	169	196
Safety Follow-Up[b]	Date of last dose + 7 days	Date of last dose + 30 days	Date of last dose + 60 days

[a] Start day of Week 4 for study day 1 includes only post-Baseline assessments that occurred after the first dose.
[b] The safety follow-up visit window will be restricted to assessments prior to the date of initiation of another investigational agent or hormonal therapy affecting the hypothalamic-pituitary-gonadal axis or surgical intervention for uterine fibroids.

Table 3: Visit Windows for Week 12/Week 24 Assessments (ECG, BMD, UFS-QoL)

Visit	Start Day	Target Day	End Day
Week 12	64	85	106
Week 24	148	169	196
Safety Follow-up[a]	Date of last dose + 7 days	Date of last dose + 30 days	Date of last dose + 60 days

Abbreviations: BMD, bone mineral density; ECG, electrocardiogram; UFS-QoL, Uterine Fibroid Symptom and Health-Related Quality of Life.

[a] The safety follow-up visit window will be restricted to assessments prior to the date of initiation of another investigational agent or hormonal therapy affecting the hypothalamic-pituitary-gonadal axis or surgical intervention for uterine fibroids

Table 4: Visit Windows for Week 24 Assessments (Transvaginal Ultrasound, Endometrial Biopsy, EQ-5D-5L)

Visit	Start Day	Target Day	End Day
Week 24	128	169	196
Safety Follow-up[a]	Date of last dose + 7 days	Date of last dose + 30 days	Date of last dose + 60 days

Abbreviations: EQ-5D-5L, European Quality of Life Five-Domain Five-Level.

[a] The safety follow-up visit window will be restricted to assessments prior to the date of initiation of another investigational agent or hormonal therapy affecting the hypothalamic-pituitary-gonadal axis or surgical intervention for uterine fibroids.

Table 5: Time Window for eDiary and Feminine Product Collection

Visit	Feminine Product Collection Visit Date[a,b]	Time Window[a]
Week 4	X_1	(Date of Study Day 1) - < X_1
Week 8	X_2	(X_1 +1) - ≤ X_2
Week 12	X_3	(X_2 +1) - ≤ X_3
Week 16	X_4	(X_3 +1) - ≤ X_4
Week 20	X_5	(X_4 +1) - ≤ X_5
Week 24	X_6	(X_5 +1) - ≤ X_6
Week 24/EOT	X_{Last}[c]	(Previous Feminine Product Returned Visit +1)] – ≤ X_{Last}

[a] If feminine products are collected at more than 1 visit within a given visit window (Table 2), the last feminine product collection date will be used to define the time window. If the patient missed the previous visit, a planned study visit date will be used to calculate the window.

[b] In the absence of feminine product collection due to amenorrhea the visit date when amenorrhea was reported will be used.

[c] Date of last non-missing feminine product collection within the interval from (last dose date – 35) to (last dose date + 7 days) (see Section 7.3.3).

4.4. General Rules for Missing Data

Handling of missing data for the primary efficacy analysis is described in Section 7.3.5.

4.4.1. By-Visit Endpoints

By-visit endpoints will be analyzed using observed data, unless otherwise specified. For observed data analyses, missing data will not be imputed and only the observed records will be included.

4.4.2. Adverse Events and Concomitant Medications

The following imputation rules for the safety analyses will be used to address the issues with partial dates. The imputed dates will be used to determine the treatment-emergent period. For adverse events with a partial date, available date parts (year, month, and day) of the partial date will be compared with the corresponding date components of the start date and end dates of the treatment-emergent period to determine if the event is treatment emergent. When in doubt, the adverse event will be considered treatment emergent by default.

The following rules will be applied to impute partial dates for adverse events:

- If start date of an adverse event is partially missing, impute as follows:
  - If both Month and Day are missing and Year = Year of treatment start date, then set to treatment start date as long as adverse event end date is not prior to treatment start date;
  - If both Month and Day are missing and Year ≠ Year of treatment start date, then set to January 1;
  - If Day is missing and Month and Year = Month and Year of treatment start date, then set to treatment start date as long as adverse event end date is not prior to treatment start date;
  - If Day is missing and Month and Year ≠ Month and Year of treatment start date, then set to first of the month;
  - If start date is completely missing, set to treatment start date as long as adverse event end date is not prior to treatment start date.
- If end date of an adverse event is partially missing, impute as follows:
  - If both Month and Day are missing, then set to December 31;
  - If only Day is missing, then set to last day of the month;
  - If end date is completely missing, do not impute.

When the start date or end date of a medication is partially missing, the date will be imputed to determine whether the medication is prior or concomitant (or both).

The following rules will be applied to impute partial dates for medications:

- If start date of a medication is partially missing, impute as follows:
  - If both Month and Day are missing, then set to January 1;

  - If only Day is missing, then set to the first of the month.
- If end date of a medication is partially missing, impute as follows:
  - If both Month and Day are missing, then set to December 31;
  - If only Day is missing, then set to last day of the month.

If start date or end date of a medication is completely missing, do not impute.

5. STUDY POPULATION

5.1. Subjects Disposition

The number of patients for each of the following categories will be summarized by treatment group:

- All randomized patients;
- Patients included in the Safety population;
- Patients who completed the 12-Week randomized treatment period;
- Patients who completed the 24-Week randomized treatment period;
- Patients who discontinued early from the 24-Week randomized treatment period and reasons for discontinuation;
- Patients who enrolled in the extension study;
- Patients who entered the Post-Treatment Follow-Up Period and did not enroll in the extension study.

Patient disposition will be summarized for all randomized patients. Summaries will include the number and percentage of patients in the mITT and Safety populations. The number and percentage of patients who prematurely discontinue study drug and the reasons for discontinuation will be summarized by treatment group. The number and percentage of patients who continue into the extension study (MVT-601-3003) will also be summarized by treatment group.

5.2. Screen Failure

Reasons for screen failure will be summarized. Number and percentage of patients who did not pass screening will be based on the patients who signed the informed consent form but were not randomized.

5.3. Protocol Deviations

Protocol deviations will be categorized as important or minor per the protocol deviation plan. Important protocol deviations will include, but will not be limited to, the following categories:

- Randomized patient who did not satisfy key entry criteria;
- Randomized patient who met withdrawal criteria during the study but was not withdrawn;
- Randomized patient who received the wrong treatment;
- Randomized patient who received a prohibited concomitant medication that met criteria for an important protocol deviation;
- Unintentional unblinding of treatment assignment.

Important protocol deviations will be summarized by deviation category for all patients in the mITT population. A patient listing of all important protocol deviations will be provided.

In addition, patient eligibility, including inclusion criteria that are not met and exclusion criteria that are met at randomization enrollment, will be summarized for all patients in the mITT population.

A selected subset of the major protocol deviations that are likely to affect analysis of efficacy will be identified to define the Per-Protocol population prior to the database lock. This subset will include but will not be limited to the following important protocol deviations:

- Did not satisfy key entry criteria **(**restricted to patients with missing Baseline MBL volume or ineligible Baseline MBL volume);
- Drug compliance < 75%;
- Patient received prohibited concomitant medications that met criteria for important protocol deviation: restricted to patients who received prohibited concomitant medications that may cause significant drug-drug interaction;
- Unintentional unblinding of treatment assignment.

5.4. Demographic and Baseline Characteristics

Demographic and Baseline characteristics will be summarized by treatment group for the mITT population. Categorical data will be summarized using frequencies and percentages, by treatment group and overall (see Table 6 below). Summaries of continuous data will display the mean, SD, median, minimum, and maximum. The numbers of missing values will also be summarized.

Table 6: Categories for Demographic and Baseline Characteristics

Variable	Category
Age (years)	< 40, ≥ 40
Geographic region	North America, Rest of World
Race	Black or African American, White, American Indian or Alaska Native, Asian, Native Hawaiian or Other Pacific Islander, Other
Ethnicity	Hispanic or Latino, Not Hispanic or Latino or Not reported
BMI (kg/m^2) at Baseline	< 18.5, 18.5 to <25, 25 to <30, 30 to < 35, 35 to < 40, ≥ 40
History of prior pregnancy	Yes, No
Disease duration of uterine fibroid (years)	Min to <1, ≥ 1 to < 3, ≥3 to <5, ≥5 to <10, ≥ 10
Type of uterine fibroids	
Subserous fibroid	Yes, No
Intramural fibroid	Yes, No
Submucosal fibroid	Yes, No
Other	Yes, No
Any surgery for uterine fibroids	Yes, No
Volume of myoma at Baseline (cm^3)	< 25, ≥ 25
Volume of uterus at Baseline (cm^3)	< 300, ≥ 300
Menstrual blood loss volume at Baseline (mL)	< 225, ≥ 225
Menstrual blood loss volume at Baseline (mL)	< 160, ≥ 160
Hemoglobin at Baseline (g/dL)	Min to < 8, ≥ 8 to <10.5, ≥ 10.5 to <12, ≥ 12
UFS-QoL Bleeding and Pelvic Discomfort Scale	0 to < 25, 25 to <50, 50 to <75, 75 to 100
Maximum NRS score for uterine fibroid-associated pain at Baseline	< 4, ≥ 4
Patient Global Assessment	
Function	No limitation at all, mild limitation, moderate limitation, quite a bit of limitation, extreme limitation
Symptoms	Not severe, mildly severe, moderately severe, very severe, extremely severe

Abbreviations: BMI = body mass index; NRS = Numerical Rating Scale; UFS-QoL = Uterine Fibroid Symptom and Health-Related Quality of Life.

5.5. Medical History

Medical history will be coded using the Medical Dictionary for Regulatory Activities (MedDRA) and will be summarized by system organ class (SOC) and preferred term (PT). Additionally, summaries of uterine fibroid–specific medical and surgical treatment history will be provided. A patient with multiple occurrences of medical history within a PT will be counted only once in that PT.

5.6. Prior Medications and Concomitant Medications

Prior medications and concomitant medications taken during the study treatment period will be summarized for all patients in the Safety population by treatment group. Medications are considered concomitant if exposure occurs during the treatment period.

The number and percentage of patients who took at least one dose of a prior medication for treatment of uterine fibroids will be summarized by treatment group and overall using the World Health Organization (WHO) Drug Dictionary and summarized according to the Anatomical Therapeutic Chemical (ATC) Classification System and generic medication name. A patient who has been administered several medications with the same preferred medication name will be counted only once for that preferred medication name.

6. STUDY DRUG EXPOSURE AND COMPLIANCE

Patients in the Safety population will be summarized for extent of exposure and compliance to study drug by actual treatment received. Exposure to and compliance with relugolix (or relugolix placebo) and E2/NETA (or placebo) will be summarized separately and will be based on the drug accountability case report forms.

Study drug exposure summaries will include the total dosage taken in milligrams, the total number of tablets (or capsules) taken, and the treatment duration.

Study drug compliance will be summarized for the treatment period and will be calculated as follows:

(total tablets taken / total tablets expected to be taken) x 100

The total tablets taken will be calculated as:

(total tablets dispensed - total tablets returned)

The total tablets expected to be taken is calculated as the total number of tablets a patient is expected to take each day times the length of time (in days) that the patient was in the treatment period of the study. Tablets that were dispensed and not returned will be assumed to have been taken. For patients who did not return for their last scheduled visit, tablets that were dispensed and not returned will not be included in the calculation of study drug compliance. For patients who did not return for any post-Baseline visits and did not return dispensed study drug, study drug compliance will not be calculated and will be categorized as "not able to calculate" in summaries of study drug compliance.

Summary statistics of study drug compliance (eg, mean, median, etc.) will be presented, along with a categorical summary (eg, ≤ 80%, 80 to 100%, > 100%).

7. EFFICACY ANALYSES

7.1. General Considerations

Efficacy analyses will be conducted on the mITT population according to the randomized treatment assignment. Stratified analyses will incorporate the randomization stratification factors. If the group of patients at any factor level from a randomization stratification factor (eg, patients with Baseline MBL volume ≥ 225 mL) comprises < 10% of the entire mITT population, this stratification factor (eg, Baseline MBL volume) will not be used for stratified analyses. In addition, if there are < 15 patients in 1 of the 4 strata (derived from the 2 stratification factors each with 2 levels), only stratification factor of Baseline MBL volume (< 225 versus ≥225 mL) will be used in the stratified analysis for more robust strata-adjusted estimation of treatment effect. The stratification category used at the time of randomization (in the Interactive Web Recognition Service [IWRS] system) will be used for all analyses rather than data recorded on the electronic case report form (eCRF) unless otherwise specified. A sensitivity analysis of the primary endpoint will be performed if the data in the IWRS and eCRF for stratification factors differ by > 5%.

7.1.1. Analyses for Binary Data and Other Categorical Data

Binary data will be summarized by frequency counts and percentages for each treatment group.

7.1.2. Analyses for Categorical Data

Qualitative variables will be summarized by frequency counts and percentages. Unless otherwise specified, the calculation of proportions will include the missing category. Therefore, counts of missing observations will be included in the denominator and presented as a separate category.

7.1.3. Analyses for Continuous Data

Continuous variables will be summarized using descriptive statistics (eg, *n*, mean, median, SD, minimum, maximum, and first and third quartiles). For the analyses of change from Baseline, the mean at Baseline will be calculated for all patients with at least one post-Baseline value by treatment group. Additionally, the mean will also be calculated for each visit, including only the patients who are in the analysis who have data for that visit by treatment group.

7.1.4. Analyses for Time to Event Data

Time-to-event endpoints will be summarized using the Kaplan-Meier method. The median, quartiles, and probabilities of an event at particular time points will be estimated by the Kaplan-Meier method.

Confidence interval for the Kaplan-Meier estimation is calculated using the exponential Greenwood formula via log-log transformation of the survival function.

The variance of the treatment difference will be calculated using the following formula:

$$V\left[\widehat{S_R}(t) - \widehat{S_L}(t)\right] = \hat{V}\left[\widehat{S_R}(t)\right] + \hat{V}[\widehat{S_L}(t)];$$

where each of the component of the variance of the Kaplan-Meier estimate will be calculated using Greenwood's formula:

$$\hat{V}[\hat{S}(t)] = \hat{S}(t)^2 \sum_{t_i \le t} \frac{d_i}{n_i(n_i - d_i)}$$

where n_i denotes the number of patients at risk at time t_i, and d_i denotes the number of events observed at time t_i.

The 95% CI of the treatment difference will be calculated using a log-log transformation of the difference in survival function, as follows:

$$[(\widehat{S_R}(t) - \widehat{S_L}(t))^{exp(1.96\,\hat{\tau}(t))}, (\widehat{S_R}(t) - \widehat{S_L}(t))^{exp(-1.96\,\hat{\tau}(t))}]$$

$$\text{where } \hat{\tau}^2(t) = \frac{\hat{V}[\widehat{S_R}(t) - \widehat{S_L}(t)]}{\{[\widehat{S_R}(t) - \widehat{S_L}(t)]\log[\widehat{S_R}(t) - \widehat{S_L}(t)]\}^2}.$$

A stratified log-rank test will be used to compare each relugolix arm to placebo. Randomization stratification factors will be used to stratify inferential testing.

7.2. Multiplicity Adjustment

The primary and the ranked secondary efficacy analyses will be performed at an overall alpha level of 0.05 (two-sided) comparing relugolix + E2/NETA (Group A) with placebo (Group C). A test will be deemed statistically significant if the two-sided p-value rounded to four decimal places is < 0.05. A gate-keeping testing procedure will be applied to maintain the family-wise type I error rate for the testing of primary and ranked secondary endpoints (see Section 7.4.1 for details).

Comparative statistics (p-values, 95% CIs for differences) will be provided for the treatment comparison of relugolix + E2/NETA with placebo for all other secondary efficacy endpoints. A treatment comparison of relugolix + delayed E2/NETA (Group B) with placebo will be performed only for the primary efficacy endpoint. There will be no statistical testing for treatment differences between the relugolix groups (Group A versus Group B) for any efficacy endpoints. The relugolix + E2/NETA group and relugolix + delayed E2/NETA group will be compared for the following safety endpoints: percent change from Baseline to Week 12 in BMD and incidence of vasomotor symptoms by 12 weeks (see Section 9.3.5 and Section 9.1.7, respectively). The above comparative analyses are not part of the gate-keeping testing procedure for label claims. p-values for primary and key secondary endpoints were adjusted for multiplicity. All other p-values are provided at a nominal level of 0.05.

7.3. Primary Efficacy Endpoint

The primary efficacy endpoint of the study is the proportion of women who achieve an MBL volume of < 80 mL AND at least a 50% reduction from Baseline in MBL volume over the last 35 days of treatment as measured by the alkaline hematin method. The primary endpoint will be referred to as responder rate and derived on the basis of the total MBL volume measured at the Week 24/EOT visit window taking into consideration the patient's compliance with return of feminine products and completion of the eDiary (see Section 7.3.2 and Section 7.3.4 for details).

7.3.1. Primary Efficacy Analysis

The following primary hypothesis for the primary efficacy endpoint will be tested:

Null hypothesis H_{01}: $\pi_R \leq \pi_P$ versus Alternative hypothesis H_{a1}: $\pi_R > \pi_P$

where π_R and π_P are the responder rates at Week 24/EOT for relugolix + E2/NETA (Group A) and placebo (Group C), respectively.

The treatment comparison between the relugolix + E2/NETA and the placebo will be analyzed using a Cochran-Mantel-Haenszel test statistic for proportions stratified by the Baseline mean MBL volume using the alkaline hematin method (< 225 mL versus ≥ 225 mL) and geographic region (North America versus Rest of World). The difference in responder rates between the relugolix + E2/NETA and placebo and its two-sided 95% CI will be estimated using stratum-adjusted Mantel-Haenszel proportions. The unadjusted responder rates and the difference in responder rates between the relugolix + E2/NETA and placebo groups and the corresponding two-sided 95% CI also will be provided. The study will be considered positive if the treatment effect for the primary endpoint is statistically significant with two-sided p-value < 0.05.

For the primary analysis, primary endpoint will incorporate the missing data handling rules described in Section 7.3.5.

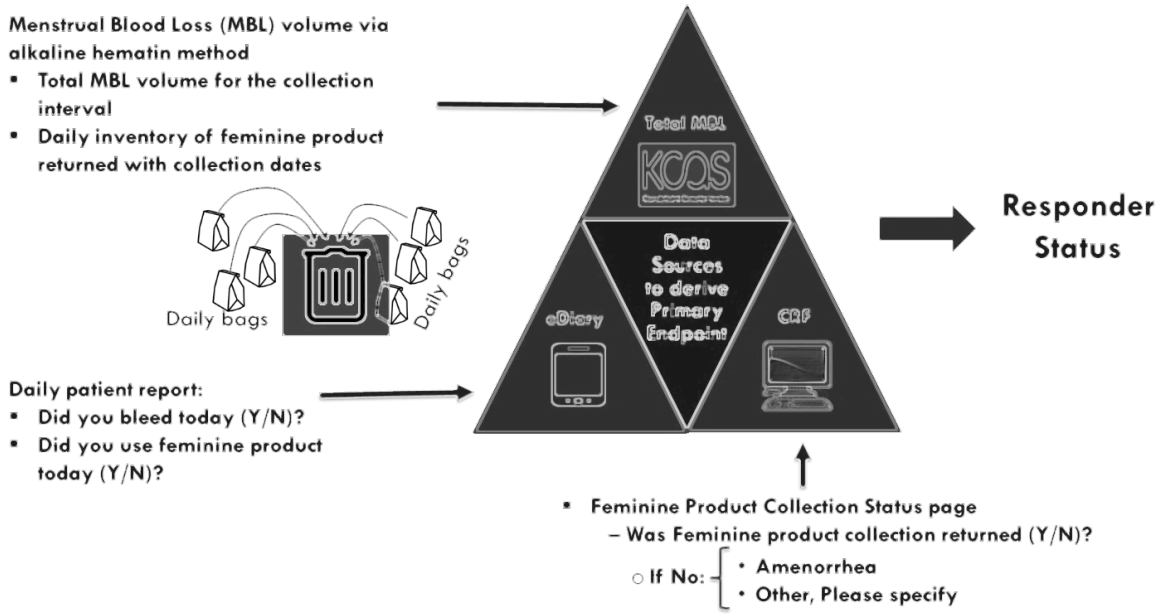
7.3.2. Data Sources Supporting Derivation of Responder Status

The data sources that will be used to support derivation of responder status, the primary endpoint, are depicted in Figure 2 below. They include:

- Menstrual blood loss volume determined by the alkaline hematin method;
- Daily patient report of bleeding (yes/no) and use of feminine product (yes/no) captured in the eDiary;
- The status of feminine product (FP) collection return (yes/no) recorded on the eCRF page at each visit with specific reasons captured when no product collection was returned.

The total MBL volume is reported from the analysis of FP returned for each collection interval. An inventory of days (with dates) for which FP was collected and returned is also available. This inventory is aligned with patients' reports of bleeding and FP use in the eDiary. The status of FP collection return, and specifically the reason for non-return of FP reported on the Feminine Product Collection eCFR page is used to support derivation of responder status (see Section 7.3.5 for details).

Figure 2: Data Sources Supporting Derivation of Primary Endpoint



Abbreviations: CRF = case report form.

7.3.3. Definitions Related to Menstrual Blood Loss

Menstrual Blood Loss Volume

All returned feminine products (validated, validated but unauthorized, or unvalidated products) collected at each visit will be analyzed by the alkaline hematin method to obtain the MBL volume. The MBL volume measured over the Week 24/EOT feminine product collection interval (up to 35 days prior to the last dose of treatment) will be used for analysis of the primary efficacy endpoint (see details below). The vendor, KCAS, reports when unauthorized feminine products (products not dispensed for use in the trial) have been returned. KCAS also reports whether the unauthorized products have previously been validated for their analysis. The report details MBL volumes for authorized, unauthorized but validated, and unauthorized and unvalidated products.

Validated Menstrual Blood Loss Volume

All returned feminine products collected at each visit, with the exception of unvalidated products, will be assessed by the alkaline hematin method to obtain the validated MBL volume. The validated MBL volume is derived from assessments of all returned validated feminine products (including validated and validated but unauthorized products) and will be used for sensitivity analysis.

Baseline Menstrual Blood Loss Volume

Baseline MBL volume is defined as the average MBL volume from the one or two consecutive screening menstrual cycles used to meet the inclusion criteria prior to the date of the first dose of study drug as assessed by the alkaline hematin method as follows:

For patients with MBL volume ≥ 160 mL during the screening period, the Baseline MBL volume is the last measurement collected before the first administration of study drug.

If the MBL volume is < 160 mL, the Baseline MBL volume is defined as the average of the MBL volume from the two screening menstrual cycles used to meet the inclusion criteria prior to the date of the first dose of study drug as assessed by the alkaline hematin method (see Figure 4-2 of the study protocol for details).

Week 24/EOT Feminine Product Collection Interval

To ensure collection of all feminine products used during that menstrual cycle, an interval of up to 35 days for measurement of the primary endpoint was selected to accommodate women who continue to have cyclic bleeding on study treatment and whose natural cycle was at the upper end of the normal cycle duration range. This method is consistent with that used during screening for collection of feminine products. Specifically, the feminine product collection interval at Week 24/EOT is driven by types of bleeding patterns experienced by the patients, as described below:

- For patients who continue to have cyclic bleeding, the length of the interval depends on the duration of the patient's natural cycle; this is consistent with the way the Baseline MBL volume was determined (eg, the interval ranging from approximately 21 to 35 days);
- Patients who report irregular, non-cyclic bleeding are instructed to collect and return all feminine product used between study visits, up to 35 days, as per the schedule of events;
- For patients who report amenorrhea on the feminine production collection eCRF page, an interval of last 35 days of treatment will be reviewed to ensure that reported amenorrhea is not due to incomplete collection.

For patients who are in the midst of an episode of cyclic bleeding at the time of the Week 24/EOT visit, the visit window may be extended up to 7 days after the last dose of study drug to ensure patients return all used feminine products over that bleeding episode.

Per protocol, all used feminine products are to be collected at each visit and returned for analysis using the alkaline hematin method. For patients who continue to have menstrual bleeding, study visits are timed such that the feminine products used in the entire menstrual bleeding cycle are collected in one container provided at each visit.

MBL Volume at Week 24/EOT

MBL volume at Week 24/EOT is defined as the MBL volume obtained from the feminine product returned over the Week 24/EOT feminine product collection interval, as described above. The MBL volume at Week 24/EOT will be used to derive the primary efficacy endpoint.

If a patient did not return feminine product over the last 35 days of treatment and reported amenorrhea on the feminine product return eCRF page, she will be considered as amenorrhoeic and her MBL volume will be assigned as 0 mL.

Feminine Product Return Rate at Week 24/EOT

To quantify degree of compliance with feminine product collection, the FPRR will be calculated based on the inventory of feminine product returned by day (dates) summarized on the Feminine Product Collection eCFR page (provided by the vendor, KCAS) and responses to the eDiary Question 4 regarding bleeding experience and Question 5 regarding the use of feminine product obtained for the corresponding eDiary window (see Table 5). Specifically:

- For those who returned feminine product at Week 24/EOT, the FPRR was calculated as the observed number of days with returned feminine products (based on the inventory of FP received by KCAS) divided by the expected number of days with bleeding and use of product as reported on the eDiary within the Week 24/EOT feminine product collection interval (as defined above).
- For those who did not return any feminine products:
  - If the reason was amenorrhea reported on the eCRF or if spotting/negligible bleeding was reported on the eCRF and confirmed by eDiary over the Week 24/EOT visit window, their FPRR will be set to 100% because the lack of menstruation obviates the need for feminine product collection.
  - Otherwise if the reason is any other, their FPRR was set to 0.

$$\text{FPRR} = \frac{\text{observed (No. of days with returned FP [per KCAS])}}{\text{expected (No. of days reported bleeding and use of FP [per eDiary])}} x100$$

Return of feminine products will be summarized in the CSR for Week 24/EOT visit.

7.3.4. Definition of Responder at Week 24/EOT

A responder at Week 24/EOT is defined as a patient who satisfies both the following:

- Had MBL volume of < 80 mL at Week 24/EOT;
- Had at least a 50% reduction from Baseline in MBL volume at Week 24/EOT.

The reduction from Baseline in MBL volume at Week 24/EOT will be calculated as the absolute change at Week 24/EOT in MBL volume from the Baseline MBL volume divided by the Baseline MBL volume.

Responder status at Week 24/EOT will be assessed based on the reported MBL volume at Week 24/EOT, in conjunction with treatment duration, compliance with feminine product collection, and compliance with eDiary entry over the same visit window (see Section 7.3.5 for details).

7.3.5. Derivation of Responder Status at Week 24/EOT and Missing Data Handling Rules

For the evaluation of primary endpoint, missing data handling rules will be implemented for deriving responder status at Week 24/EOT as described below. The following elements will be checked: duration of treatment exposure; compliance with feminine product collection against the eDiary, as measured by FPRR; compliance with eDiary entry, defined as the proportion of eDiary entry days over the length (days) of FP collection interval for Week 24/EOT visit; and reasons for no FP collection (as displayed in Table 7).

Patients with < 4 weeks of treatment who withdraw from the study prematurely due to lack of efficacy or withdraw from the study prematurely to undergo surgical intervention for uterine fibroids will be considered as non-responders.

All other patients will have their responder status determined as follows:

- For patients with a FPRR of 100%, responder status will be determined based on the observed MBL volume;
- For patients who had incomplete feminine product collection, with a FPRR of < 100%, responder status will be derived based on either imputed or observed MBL volume;
  - Those with an MBL volume ≥ 80 mL or < 50% reduction from Baseline will be considered as non-responders;
  - Those with an MBL volume < 80 mL and ≥ 50% reduction from Baseline will be imputed for partial or complete missing MBL volume (see Section 7.3.6 for details).
- For patients who did not return a feminine product collection, responder status will be determined depending on the reason reported on the Feminine Product Collection eCRF:
  - If the reason is reported as Amenorrhea, the last 35 days of treatment will be used to derive responder status:
    - If the Week 24/EOT interval was 35 days, then she will be considered as a responder;
    - If the Week 24/EOT interval was <35 days, the following supportive information will be used to derive responder status:
      - If a patient reported amenorrhea at the visit prior to Week 24/EOT, she will be defined as a responder;
      - If a patient did not report amenorrhea at the visit prior to Week 24/EOT, eDiary data from the prior visit interval will be reviewed to confirm whether the patient was amenorrheic for a total of 35 days.

- If the eDiary from the previous interval confirms amenorrhea, then the patient will be considered as a responder;
- Otherwise, MBL volume will be imputed.

- If the reason is Other and the specification describes spotting or negligible bleeding, responder status will be defined as follows:
  - The patient will be considered as a responder if it is supported by the eDiary data: the eDiary entry rate must exceed 70% and the patient must have reported no more than 5 total days of bleeding with product use and no more than 3 consecutive bleeding with product use over the collection interval.
  - If the eDiary entries did not confirm spotting or negligible bleeding, but the patient had at least 8 weeks of MBL volume data prior to the Week 24/EOT visit, her missing MBL volume will be imputed to determine responder status. Eight weeks of MBL volume data represents a reasonable minimum length of observation to justify imputation of the remaining data in assessing the effects of hormonal therapy.
  - Otherwise if the patient had < 8 weeks of MBL volume data, she will be considered as a non-responder;
- If the reason is any Other, the responder status will be derived as follows:
  - If the patient had at least 8 weeks of MBL volume data prior to the Week 24/EOT visit, her missing MBL volume will be imputed and her responder status will be based on the imputed MBL volume.
  - If the patient had < 8 weeks of MBL volume data, she will be considered as a non-responder.

Table 7: Derivation of Responder Status at Week 24/EOT and Missing Data Handling Rules – for Primary Analysis

Treatment Exposure	FP Collection (FPRR)	Observed MBL Volume	Reason for No FP Collection	Responder Status
< 4 weeks	N/A	N/A	N/A	Imputed as non-responder
≥ 4 weeks	100% FP Compliance	N/A	N/A	Based on the observed MBL volume
	<100% FP Compliance	MBL volume ≥ 80 mL or <50% reduction from Baseline	N/A	Imputed as non-responder based on the observed MBL volume
		MBL volume < 80 mL and ≥ 50% reduction from Baseline	N/A	**Based on the imputed MBL volume**
	No FP Collection	N/A	Reported "Amenorrhea"	Imputed as responder
			Reported "Spotting or negligible bleeding" and confirmed by eDiary[a]	Imputed as responder
			Reported "Spotting or negligible bleeding" although not confirmed by eDiary or any other reason, had at least 8 weeks of MBL volume data	**Based on the imputed MBL volume**
			The entries in the eDiary did not verify "Spotting or negligible bleeding" or any other reason and if had < 8 weeks of MBL volume data	Imputed as non-responders

Abbreviations: eDiary, electronic diary; FP, feminine product; EOT, end of treatment; MBL, menstrual blood loss; N/A, not available.

[a] Defined as those patients who meet the following criteria: eDiary entry rate > 70% and no more than 3 consecutive days and no more than 5 days of bleeding/spotting and use of feminine product reported on the eDiary over the Week 24/EOT visit window (see Table 5).

7.3.6. Mixed-Effects Model for Imputing Missing or Partially Missing MBL Volume at Week 24/EOT

For the primary analysis, patients with missing MBL volumes at Week 24/EOT will be identified per missing data handling rules as described above. For imputing missing data for the primary analysis, a mixed-effects model approach will be used, as the mixed-effects approach may better describe the effects of a hormonal treatment (such as suppression of the hypothalamic-pituitary-ovarian axis by GnRH antagonists).

Specifically, a mixed-effects model with repeated measures of MBL volumes at multiple time points (Weeks 4, 8, 12, 16, 20 and 24) will be fitted to predict percent change in MBL volume from Baseline (as a dependent variable) through the fixed-effects associated with covariates (ie, stratification factors of Baseline MBL volume and geographic region, visit, treatment, and visit by treatment interaction) and random effects (from the individual patients). In this model, an unstructured variance-covariance matrix is assumed for each patient.

See sample SAS codes below for illustration where PCHG_MBL is percent change in MBL volume from Baseline as a dependent variable, PID is patient identification number, BMBL is a randomization stratification factor (Baseline MBL < 225 vs ≥ 225), REGION is a randomization stratification factor (North America vs Rest of World), TRT is treatment group (relugolix + E2/NETA or Placebo), VISIT is visit time point (4, 8, 12, 16, 20, and 24 weeks) and TRT*VISIT is the visit by treatment interaction. The specification of type=UN implements unstructured variance-covariance matrix for an individual patient with multiple measures of MBL volumes.

```
proc mixed data=MBL_dataset method=REML covtest;
class PID BMBL REGION TRT VISIT;
model PCHG_MBL= BMBL REGION VISIT TRT VISIT*TRT/s outp=ufmi_mixed_p
covb;
repeated VISIT /type=UN subject=PID r;
lsmeans TRT/diff;
ods output SolutionF=mixparms CovB=mixcovb;
```

Applying this model over the observed longitudinal MBL volume data, the fixed-effects will be estimated and relationship of percent change in MBL volume from Baseline with the covariates will be characterized by the fitted model. From the fitted model, the percent change in MBL volume (whether missing or not) will be predicted for each patient at each visit and in a particular stratum. The imputed MBL volume will be obtained by first multiplying the imputed percent change with the individual patient's Baseline MBL volume to the difference, and then adding the Baseline BML volume to the difference.

The main reason for using percent change in MBL volume over reported MBL volume as a dependent variable in the mixed-effects model is that the percent change is part of the derivation of the primary endpoint. Secondly, the percent change is a normalized value adjusted for the Baseline value and less influenced by Baseline MBL volume, and therefore it is a better metric to describe the relationship of MBL volume reduction with hormonal treatment and to impute the missing volumes in a more robust fashion.

Since the purpose of using a mixed-effects model is imputing the missing MBL volumes identified at Week 24/EOT, the predicted MBL volumes at the corresponding Week 24/EOT visit will be used to determine responder status. For patients without the need for imputation, their responder status will be derived according to the algorithms laid out in Table 7. This imputation approach is consistent with the definition of responder at Week 24/EOT for the primary analysis.

7.3.7. Sensitivity Analyses

To assess the robustness of the primary analysis, the following sensitivity analyses of the primary endpoint will be conducted at Week 24/EOT.

7.3.7.1. Sensitivity Analysis 1

To assess the potential impact of unvalidated feminine product use, the primary endpoint will be analyzed as sensitivity analysis in a similar fashion to the primary analysis using the Week 24/EOT validated MBL volume (obtained from the validated or validated-but-unauthorized feminine products only and excluding unvalidated products).

7.3.7.2. Sensitivity Analysis 2

To assess the potential impact of missing data due to inadequate collection of feminine products, the primary endpoint will be analyzed with a sensitivity analysis using the missing data handling rules as described in Table 8 below where the observed MBL volume will be used to assess the responder status at Week 24/EOT when feminine product collection was incomplete. These rules differ from those used in the primary analysis in that no imputation will be implemented for patients with < 100% feminine product compliance and the reported MBL volume both < 80 mL and a ≥ 50% reduction from Baseline as highlighted in Table 8.

Table 8: Derivation of Responder Status at Week 24/EOT and Missing Data Handling Rules – for Sensitivity Analysis

Treatment Exposure	FP Collection (FPRR)	Observed MBL Volume	Reason for No FP Collection	Responder Status
< 4 weeks	N/A	N/A	N/A	Imputed as non-responder
≥ 4 weeks	100% FP Compliance	N/A	N/A	Based on the observed MBL volume
	< 100% FP Compliance	MBL volume ≥ 80mL or < 50% reduction from Baseline	N/A	Imputed as non-responder based on the observed MBL volume
		MBL volume < 80mL and ≥ 50% reduction from Baseline	N/A	**Based on the observed MBL volume**
	No FP Collection	N/A	Reported "Amenorrhea"	Imputed as responder
			Reported "Spotting or negligible bleeding" and confirmed by eDiary[a]	Imputed as responder
			Reported "Spotting or negligible bleeding" although not confirmed by eDiary or any other reason, had at least 8 weeks of MBL volume data	Based on the imputed MBL volume
			The entries in the eDiary did not verify "Spotting or negligible bleeding" or any other reason and if had < 8 weeks of MBL volume data	Imputed as non-responders

Abbreviations: eDiary, electronic diary; FP, feminine product; EOT, end of treatment; MBL, menstrual blood loss; N/A, not available.

[a] Defined as those patients who meet the following criteria: eDiary entry rate >70% and no more than 3 consecutive days and no more than 5 days of bleeding/spotting and use of feminine product reported on the eDiary over the Week 24/EOT visit window (see Table 5).

7.3.7.3. Sensitivity Analysis 3

To assess the potential impact of early discontinuation on the primary endpoint, the primary endpoint will be analyzed with a sensitivity analysis defining the patients' responder status as follows:

- Patients who discontinued study drug during the first 4 weeks for any reason or who discontinued study drug between Week 4 and Week 12 due to an adverse event, surgery or other intervention for heavy menstrual bleeding, reported lack of efficacy, or bleeding complaints will be considered as non-responders;
- All other patients will have their responder status defined using data from the Week 24/EOT assessment period using the last observation carried forward method.

7.3.7.4. Sensitivity Analysis 4

To assess the potential impact of the length and full exposure of the treatment, the primary endpoint will be analyzed for the Completers population as a sensitivity analysis. The Completers population is defined as patients in the mITT population who completed 24 weeks of study treatment.

7.3.7.5. Sensitivity Analysis 5

The primary endpoint will be analyzed on the Per-Protocol population as a sensitivity analysis, using the methods specified for the primary analysis (see definition of Per-Protocol population in Section 4.2.2).

7.3.7.6. Sensitivity Analysis 6

As a sensitivity analysis to the primary analysis using the mixed-effects model for imputing missing MBL volumes at Week 24/EOT, multiple imputation approach will be implemented as described below.

A multiple imputation method (Rubin, 1987; von Hippel, 2018) will be used to impute missing or partially missing MBL volume identified by the missing data handling rules (see Table 7 and Table 8) at Week 24/EOT as described in the following 5 steps. In this method, an arbitrary missing pattern will be assumed using Markov Chain Monte Carlo imputation to generate a monotone missing pattern for the observed longitudinal MBL volume values (including 0 mL if the patient has amenorrhea). Imputation will be performed separately by randomized treatment group (Sullivan, 2018), given the distinct bleeding patterns among the three treatment groups.

Normalizing transformations will be applied to the statistics estimated from each imputed dataset before the Rubin's combination rules can be applied (Ratitch, 2013). This combined estimation and statistical test will account for the additional variability due to imputation to provide a robust assessment of the treatment effect.

Step 1: Identifying patients with missing or incomplete MBL volume from the longitudinal MBL volume dataset as collected.

Step 2: Generating a monotone missing pattern using the Markov Chain Monte Carlo technique by imputing missing MBL volume measurements that are between non-missing results.

Step 3: Imputing the remaining missing values m = 100 times using a regression model; therefore, generating 100 complete longitudinal MBL volume datasets.

Note: if a patient missed Week 8 and prematurely discontinued study drug (eg, at Week 20) and MBL volume at Week 20 is missing or partially missing, MBL volume will be imputed for intermittent missing data at Week 8, Week 20 (EOT), and Week 24 due to discontinuation.

Step 4: Performing the same CMH test pre-specified for the primary endpoint analysis and estimating the responder rates for each arm using each of the 100 datasets based upon the MBL volume at Week 24/EOT.

Note: in the example above, the imputed MBL volume at Week 20 (EOT) will be used in the analysis, although MBL volume is imputed at Week 24.

Step 5: Combining the results from the 100 complete datasets to make inferences about the treatment effect on the responder rate.

7.3.8. Subgroup Analyses

Subgroup analyses of the primary efficacy endpoint comparing the relugolix + E2/NETA group versus the placebo group will be performed to assess whether treatment effects are consistent across clinically important subgroups. The odds ratio and its 95% CI based on a logistic regression model will be displayed in a forest plot for each subgroup. The logistic regression model will include treatment group, Baseline MBL volume value and geographic region as covariates. Subgroups will include, but will not be limited to, the subgroups outlined in Table 9.

Table 9: Planned Subgroup Analyses

Subgroup Name	Subgroup Level
Geographic region	North America vs Rest of World
Menstrual blood loss volume at Baseline (mL)	< 225 vs ≥ 225 < 120, 120 to < 160, 160 to < 225, ≥ 225
Age category (years)	< 40 vs ≥ 40 < 35, 35 to < 40, 40 to < 45, ≥ 45
Race	Black or African American vs Not Black or African American; Black or African American, White, Other
Volume of myoma at Baseline (cm^3)	< 25 vs ≥ 25
Volume of uterus at Baseline (cm^3)	< 300 vs ≥ 300
BMI (kg/m^2) at Baseline	< 30 vs ≥ 30 < 25, 25 to < 30, 30 to < 35, 35 to <40, ≥ 40
Maximum NRS score for uterine fibroid–associated pain at Baseline	< 4 vs ≥ 4
History of prior pregnancy	Yes/No

Abbreviations: BMI = body mass index; NRS = Numerical Rating Scale.

7.4. Secondary Efficacy Endpoints

Secondary efficacy variables include seven key secondary endpoints with alpha-protection and other secondary endpoints. All secondary efficacy endpoints and analyses are summarized in Appendix 1.

The treatment effect of relugolix + E2/NETA (Group A) compared to placebo (Group C) will be tested for the alpha-protected secondary endpoints using a gate-keeping procedure (see Section 7.4.1).

Comparative statistics (p-values, 95% CIs for differences) will be provided for treatment comparison of the relugolix + E2/NETA group with the placebo group for all other secondary efficacy endpoints. Treatment difference between the relugolix + delayed E2/NETA group and the placebo group will be formally tested only for the primary efficacy endpoint. There will be no statistical testing for treatment differences between the relugolix groups (relugolix + E2/NETA group against relugolix + delayed E2/NETA group) for any efficacy endpoint (see Section 7.4.2).

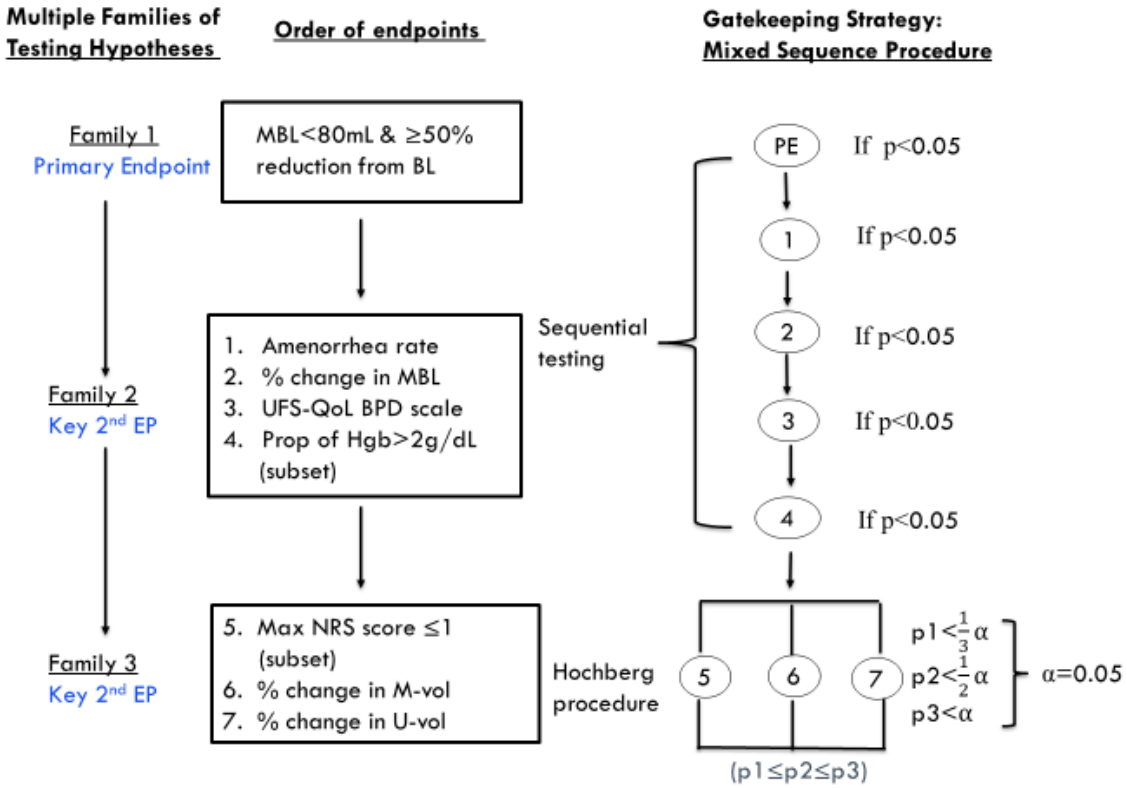
7.4.1. Key Secondary Efficacy Endpoints with Alpha-Protection

For testing whether relugolix + E2/NETA (Group A) is statistically significantly superior to placebo (Group C) for the primary efficacy endpoint as well as the seven key secondary endpoints listed below, a gate-keeping mixed sequence testing procedure will be applied to maintain the family-wise type I error rate. Under this testing procedure, the primary endpoint

will be tested first at a 2-sided 0.05 significance level. If the p-value for primary endpoint is < 0.05, the seven key endpoints listed below will be tested sequentially in the order depicted in Figure 3 (LIBERTY 1) and Figure 4 (LIBERTY 2).

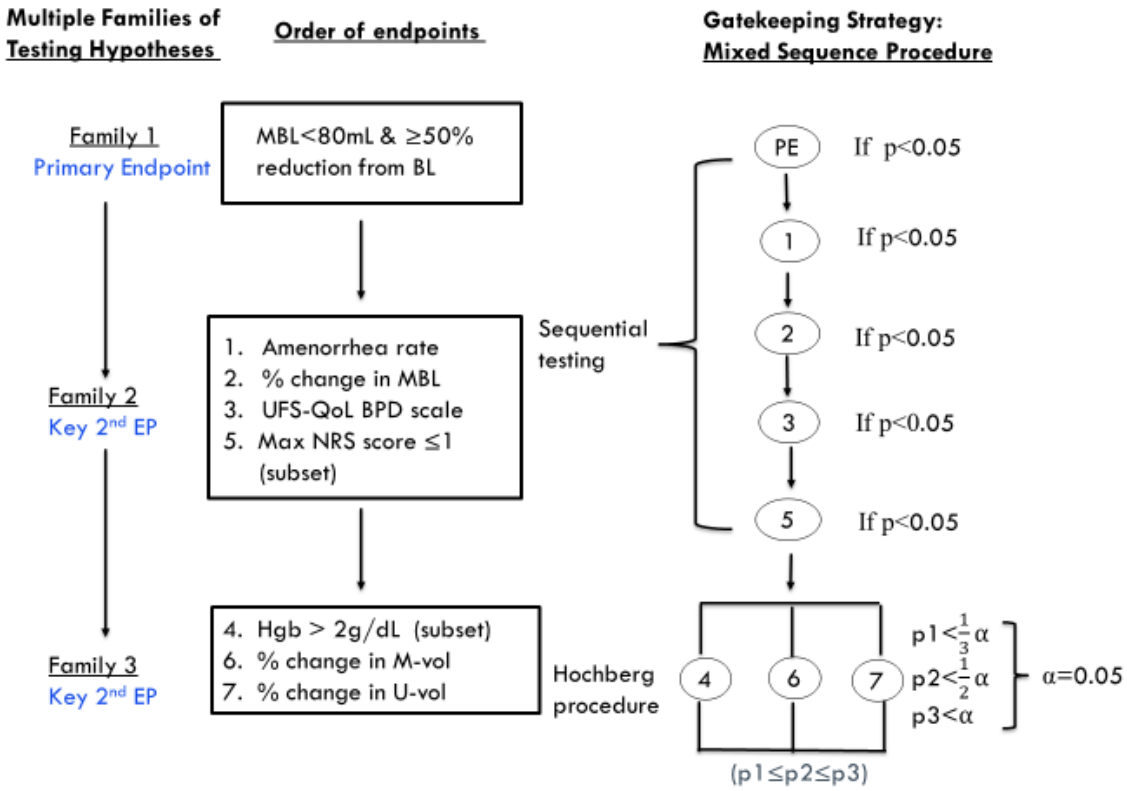
For the relugolix + E2/NETA group to be considered statistically superior to the placebo group on a secondary endpoint, the two-sided p-value must be < 0.05 for that secondary endpoint and for all higher-ranking secondary endpoints, as well as for the primary endpoint. If the two-sided p-value is < 0.05 for the fourth endpoint (proportion of women with hemoglobin ≤ 10.5 g/dL at Baseline who achieve an increase of > 2 g/dL from Baseline at Week 24 for LIBERTY 1; proportion of women who achieve a maximum NRS score ≤ 1 for uterine fibroid-associated pain over the last 35 days of treatment in the subset of women with a maximum pain score ≥ 4 during the 35 days prior to randomization for LIBERTY2), the remaining three endpoints (the fifth, sixth, or seventh) will be tested using the Hochberg step-up procedure.

Figure 3: Mixed Sequence Testing Procedure for Primary and Key Secondary Endpoints in LIBERTY 1



Abbreviations: BPD = Bleeding and Pelvic Discomfort; EP = endpoint; Hgb = hemoglobin; max = maximum; MBL = menstrual blood loss; M-vol = myoma volume; NRS = Numerical Rating Scale; PE = primary endpoint; Prop = proportion; UFS-QoL BPD = Uterine Fibroid Symptom and Health-Related Quality of Life Bleeding and Pelvic Discomfort; U-vol = uterine volume.

Figure 4: Mixed Sequence Testing Procedure for Primary and Key Secondary Endpoints in LIBERTY 2



Abbreviations: BPD = Bleeding and Pelvic Discomfort; EP = endpoint; Hgb = hemoglobin; max = maximum; MBL = menstrual blood loss; M-vol = myoma volume; NRS = Numerical Rating Scale; PE = primary endpoint; Prop = proportion; UFS-QoL BPD = Uterine Fibroid Symptom and Health-Related Quality of Life Bleeding and Pelvic Discomfort; U-vol = uterine volume.

From the Hochberg procedure, the p-values will be calculated for the three endpoints (5, 6, and 7) and ranked from the smallest to the largest. The endpoint corresponding to the largest p-value gets tested first. If the p-value is < 0.05, then no further testing will occur, and it will be concluded that all three endpoints are positive. Otherwise, the endpoint corresponding to the second largest p-value will be tested. If the p-value is < 0.025, then no further testing will occur, and it will be concluded that the endpoints corresponding to the middle and smallest p-values are positive. Otherwise, the endpoint with the smallest p-value will be tested. If the p-value is < 0.0167, no further testing will occur, and it will be concluded that only the endpoint with the smallest p-value is positive. Otherwise, all three endpoints did not pass the statistical significance criterion at 0.05 level.

The seven key secondary efficacy endpoints are as follows:

1. Proportion of women who achieve amenorrhea over the last 35 days of treatment;
2. Percent change from Baseline to Week 24 in MBL volume;

3. *Change from Baseline to Week 24 in Bleeding and Pelvic Discomfort Scale score as measured by the UFS-QoL Symptom Severity Scale (Q1, Q2, Q5);*
4. *Proportion of women with a hemoglobin ≤ 10.5 g/dL at Baseline who achieve an increase of > 2 g/dL from Baseline at Week 24*
5. *Proportion of women who achieve a maximum NRS score ≤ 1 for uterine fibroid associated pain over the last 35 days of treatment in the subset of women with a maximum pain score ≥ 4 during the 35 days prior to randomization;*
6. Percent change from Baseline to Week 24 in uterine fibroid volume;
7. Percent change from Baseline to Week 24 in uterine volume.

For key secondary efficacy endpoints (1, 4, and 5) that are evaluating proportions, treatment comparisons will be performed using a stratified Cochran-Mantel-Haenszel test with the randomization stratification factors as strata. Point estimates and 2-sided 95% CIs for treatment differences in proportions will be provided.

For key secondary endpoint 4, an increase in hemoglobin of 2g/dL is considered clinically meaningful, because it corresponds to approximately the same increase as that expected after a transfusion of ~ 2 units of packed red blood cells (Man, 2016; Bachowski, 2017).

For deriving the key secondary endpoint 5 (proportion of women who achieve a maximum NRS score ≤ 1 for uterine fibroid-associated pain over the last 35 days of treatment in the subset of women with a maximum pain score ≥ 4 during the 35 days prior to randomization), the patient subset and Week 24/EOT maximum value are determined as follows.

Because patients were asked to begin eDiary entries after returning the first collection of feminine products, the number of eDiary entries made during screening varies with the duration of screening for each patient. Some patients required only one collection to be randomized, whereas others required as many as four collections to confirm eligibility.

Once the qualifying menstruation was completed and the patient qualified for randomization based upon resulting MBL volume(s), the recording of patient's NRS scores for screening phase will be ended and the number of pain score days at Baseline can be as short as 7 days or as long as 70 days prior to randomization. If a patient meets the subset definition (maximum NRS score ≥ 4 at Baseline) over a portion of the screening days (eg, 7-70 days), she will also meet the subset definition on the entire 35 days interval.

Since the maximum NRS value is used to determine inclusion into the subset rather than an average NRS value, the variable number of days for inclusion of patients has no major impact on determining patient subset. To ensure robust estimate of response, the minimum number of non-missing daily pain scores required to calculate the maximum score at Week 24/EOT is at least 28 days (80% of the last 35 days of treatment) of pain scores recorded in the e-Diary entry.

The primary analysis of key secondary endpoint 5 will be analyzed for the subset of women who have a maximum pain score ≥ 4 during the 35 days prior to randomization and who have at least 28 days (80% of the last 35 days of treatment) of pain scores recorded in the e-Diary at Week 24/EOT. In addition, a sensitivity analysis will be conducted on the subset of women who have

a maximum pain score ≥ 4 during the 35 days prior to randomization without restricting number of days of pain scores recorded in the e-Diary.

The analysis for endpoint 5 (proportion of women who achieve a maximum NRS score ≤ 1 for uterine fibroid-associated pain over the last 35 days of treatment in the subset of women with a maximum pain score ≥ 4 during the 35 days prior to randomization) will also be performed using NRS scores reported on eDiary during menstrual and non-menstrual days.

For key secondary efficacy endpoints (6 and 7) evaluating percent change from Baseline in uterine fibroid volume and uterine volume that are measured only at Week 24, an analysis of covariance (ANCOVA) model will be used to assess treatment effect with treatment, randomization stratification factors and Baseline value as covariates.

For key secondary efficacy endpoints (2 and 3) evaluating the change (absolute or % change) from Baseline to Week 24, treatment comparisons will be performed using a mixed model repeated measures approach with treatment, visit, randomization stratification factors and treatment by visit interactions included as fixed effects and random effects (from the individual patients). In this model, an unstructured variance-covariance matrix is assumed for each patient.

7.4.2. Other Secondary Efficacy and Exploratory Endpoints

The following describes the analysis methods for other secondary efficacy endpoints and exploratory endpoints. There are three types of analyses corresponding to the three types of endpoints (time-to-event, continuous and binary) (see Appendix 1 for details).

Time-to-Event Endpoint

For time to achieving an MBL volume of < 80 mL AND at least a 50% reduction from Baseline MBL volume (as an event), time to event will be defined as weeks from date of first dose of study drug to response (event) based on the MBL volume as assessed by the alkaline hematin method. The missing data handling rules described in Section 7.3.5 for deriving responder status at Week 24/EOT will be applied similarly at Weeks 8, 12, 16, and 20. Patients without an event will be censored at the last assessment date prior to the last dose of the study drug. Kaplan-Meier methods will be used to describe the time to event distributions. A log-rank test stratified by the randomization stratification factors using the proportional hazard model (p-value from score test) will be used to compare relugolix + E2/NETA to placebo. Randomization stratification factors will be used to stratify inferential testing.

Continuous Endpoints

For endpoints evaluating the change (absolute or percent change) from Baseline to Week 24, treatment comparisons will be performed using a mixed model repeated measures approach with treatment, randomization stratification factors, visit, and treatment by visit interactions included as fixed effects. The Baseline value will be included as a covariate, and an unstructured variance-covariance matrix will be assumed. Calculation of the dependent variable (change from Baseline) for each patient at each visit will be calculated based on the visit windows specified in Section 4.3.5. Based on this model, the least squares mean at Week 24 will be compared between treatment groups and summarized along with the corresponding 95% CIs for treatment difference. In addition, summary statistics (mean change or mean % change) will be graphically presented as appropriate.

Binary Endpoints

For endpoints evaluating proportions, treatment comparisons will be performed using a stratified Cochran-Mantel-Haenszel test as appropriate with the randomization stratification factors as strata. Point estimates and 2-sided 95% CIs for treatment differences in proportions will be provided.

Descriptive statistics (point estimates and corresponding 95% CIs) will be provided by treatment group and visit as appropriate for all secondary endpoints.

Responder rate by visit (at Week 4, Week 8, Week 12, Week 16, and Week 20) will be derived in a similar fashion to the derivation of responder rate at Week 24/EOT. The missing data handling rules described in Section 7.3.5 for deriving responder status at Week 24/EOT will be applied similarly at Weeks 4, 8, 12, 16, and 20.

7.4.3. Derivation of Amenorrhea-Related Endpoints

Determination of Amenorrhea

Rules for determining amenorrhea in the treatment period is defined as those who meet 1 of the following requirements for 2 consecutive visits (approximately 56 consecutive days). Patients will be deemed to have amenorrhea during a visit window according to the following rules:

- No feminine product returned due to reported amenorrhea in 2 consecutive visits

OR

- No feminine product returned due to other reasons or feminine product collection with a negligible observed MBL volume coupled with other data indicating infrequent non-cyclic bleeding/spotting as described in Table 10.

Missing responses for menstrual bleeding questions in the eDiary will be treated as “No Bleeding” if eDiary compliance rate is > 70%.

Table 10: Rules for Determining Amenorrhea by Visit

Feminine Product Collection (KCAS) [a]	Supporting Data	
	Menstruation Status eCRF	**eDiary**
No feminine product collection due to reported amenorrhea	No menses start/stop dates reported	N/A
No feminine product collection due to other reasons	Per instructions for non-cyclic bleeding patterns, menses start date is reported but no menses stop date reported	• Data indicating infrequent, non-cyclic bleeding/spotting defined as bleeding/spotting with feminine product use for no more than 3 consecutive days and no more than 5 days bleeding total per visit window • eDiary entry rate > 70%
Feminine product collection with negligible observed MBL volume defined as <5 mL	Full or partial menses start and stop dates	• Data indicating infrequent, non-cyclic bleeding/spotting defined as bleeding/spotting with feminine product use for no more than 3 consecutive days and no more than 5 days bleeding total per visit window • eDiary entry rate > 70%

Abbreviations: eCRF, electronic case report form; eDiary, electronic diary; MBL, menstrual blood loss; N/A = not applicable.
[a] There is no requirement for feminine product return rate, as the determination of amenorrhea is based on the eDiary response.

Amenorrhea During the Last 35 Days of Treatment

Patients with amenorrhea over the last 35 days of treatment are defined as those who meet the definition of amenorrhea. A patient's amenorrhea status will also be summarized at Weeks 8, 12, 16, and 20. If a patient does not return for her Week 24/EOT visit, the eDiary responses for the last 35 days of treatment will be evaluated. If the criteria for infrequent, non-cyclic bleeding or spotting as indicated in Table 10 is met and the criteria for amenorrhea is met at the prior visit, the patient will be categorized as amenorrheic at Week 24/EOT. At all other timepoints, patients who do not return for a specific visit will be assigned as not amenorrheic at that visit.

Time to Amenorrhea

Time to amenorrhea is defined as the weeks from date of first dose of study drug to the start date of the amenorrhea window. Time to sustained amenorrhea will also be estimated and plotted using the Kaplan-Meier method.

The start date of amenorrhea is defined as the last feminine product collection date prior to start of amenorrhea. For example, if a patient's feminine product was collected at her Week 4 visit and MBL volume for this cycle did not indicate amenorrhea, and the patient reported amenorrhea on Week 8 and 12 visits, then time to start amenorrhea will be defined as starting on the date of feminine product collection for Week 4. Patients who are determined to have amenorrhea at Week 4 and Week 8 will use their Week 4 feminine product collection date as start date of amenorrhea. Patients without an event will be censored at the last assessment date prior to the last dose of the study drug.

Sustained Amenorrhea Rate by Visit

A patient's sustained amenorrhea status will be summarized at Weeks 8, 12, 16, 20, and 24, based on her time to achieving and maintaining amenorrhea until the date of last study drug dose as shown in Table 11. For example, at Week 8, a patient is considered to have achieved sustained amenorrhea status if her amenorrhea started before Week 8 and was observed every visit thereafter until the last dose of the study treatment. The proportion of patients with sustained amenorrhea will be summarized by visit. If a patient met the criteria for sustained amenorrhea but discontinues from the study, this subject's amenorrhea status will be carried forward to the Week 24 visit.

Table 11: Sustained Amenorrhea Rate by Visit

Time Point	**Amenorrhea Window**	
	Start	End
Week 8	Determined amenorrhea at Week 4	Amenorrhea observed at Week 8 and was observed at every visit thereafter until and including the last dose of study treatment
Week 12	Determined amenorrhea at Week 8	Amenorrhea observed at Week 12 and was observed at every visit thereafter until and including the last dose of study treatment
Week 16	Determined amenorrhea at Week 12	Amenorrhea observed at Week 16 and was observed at every visit thereafter until and including the last dose of study treatment
Week 20	Determined amenorrhea at Week 16	Amenorrhea observed at Week 20 and was observed at every visit thereafter until and including the last dose of study treatment
Week 24	Determined amenorrhea at Week 20	Amenorrhea observed at Week 24

7.4.4. Derivation of Patient-Reported Outcome

7.4.4.1. Numerical Rating Scale Score for Pain Associated with Uterine Fibroids

Patients completed daily eDiaries including assessment of uterine fibroid-associated pain by the Numerical Rating Scale (NRS). Patients rated their worst pain in the last 24 hours caused by their uterine fibroids on a scale from 0 to 10, with 0 indicating no pain and 10 indicating pain as bad as you can imagine. The maximum NRS score for pain at Week 24/EOT is calculated as the maximum NRS score during the last 35 days on study treatment. If any NRS scores for pain during the last 35 days on study treatment are missing, the maximum score will be calculated as the maximum of all non-missing scores. Baseline NRS score for uterine fibroid-associated pain is defined as the maximum NRS score from the 35 days of data collected prior to randomization.

Proportion of women who achieve a *maximum* NRS score for pain associated with uterine fibroids over the last 35 days of treatment that is at least a 30% reduction from Baseline will be summarized in the subset of women with a maximum pain score ≥ 4 during the 35 days prior to randomization (subset). In addition, for the subset, mean maximum NRS scores will be provided by treatment and visit. Maximum NRS score for each patient at a visit is defined as the highest NRS score reported in the visit window specified in Table 2.

7.4.4.2. UFS-QoL Score

Calculation of UFS-QoL Symptom Severity Scale Score

To calculate the Symptom Severity Scale score, a summed score is created for the items listed below and then the formula below the table is used to transform raw scores to a normalized score with a range of possible values from 0 to 100. This provides Symptom Severity Scale scores, where higher scores are indicative of greater symptom severity and lower scores indicate lower symptom severity.

Scale	Sum Item Values	Lowest and Highest Possible Raw Scores	Possible Raw Score Range
Symptom Severity	Sum 1 – 8	8, 40	32

Formula for Transformation of Symptom Severity Raw Scores ONLY:

Transformed Score = [(Actual raw score – lowest possible raw score)/(Possible raw score range)] * 100

Calculation of UFS-QoL Bleeding and Pelvic Discomfort Scale Score

The UFS-QoL Bleeding and Pelvic Discomfort (BPD) Scale has been derived from the UFS-QoL Symptoms Scale; the derivation and validation of this new scale can be found in Appendix 3. The new scale consists of the following three symptoms proximal to uterine fibroids:

- Heavy bleeding during your menstrual period (Q1)
- Passing blood clots during your menstrual period (Q2)
- Feeling tightness or pressure in your pelvic area (Q5)

To calculate the score for the BPD Scale, a summed score of the items listed below is created and then the formula below the table is used to transform the raw score to a normalized score. This provides BPD Scale scores, where higher score values are indicative of greater symptom severity and lower scores will indicate minimal symptom severity (high scores = bad).

Sub-Scale	Sum Item Values	Lowest and Highest Possible Raw Scores	Possible Raw Score Range
Q1, Q2 and Q5	Sum 1,2,5	3, 15	12

Formula for Transformation of BPD Raw Scores ONLY:

Transformed Score = [(Actual raw score – lowest possible raw score)/(Possible raw score range)] * 100

On the basis of transformed score for BPD Scale, change from Baseline in the transformed score for BPD Scale at Week 24 will be defined as an alpha-protected key secondary endpoint comparing the relugolix + E2/NETA group with the placebo group. The proportion of patients who are responders (defined as meeting a meaningful change threshold from Baseline in the BPD Scale) at Week 24 on the transformed score for the BPD Scale will be compared between the two treatment arms (the relugolix + E2/NETA group with the placebo group) using a stratified Cochran-Mantel-Haenszel test, as appropriate. The proposed responder threshold is a 20-point change. Details in the determination of the meaningful change in the BPD Scale can be found in Appendix 4.

As a descriptive assessment on robustness of the responder analysis, a plot of the cumulative distribution function (CDF) will be provided for each treatment group to display the change from Baseline to Week 24 in the transformed score for BPD Scale on the x-axis and cumulative percentage of patients experiencing up to that change on the y-axis.

Calculation of Other UFS-QoL Scale Scores and UFS-QoL Total Score

For the other UFS-QoL scales (concern, activities, revised activities, energy/mood, control, self-conscious, and sexual function), a summed score of the items listed below is created for each individual scale. To calculate the UFS-QoL total score, the values for each individual scale are summed. Using the formula below the table, all raw scores are transformed to normalized scores. Higher scores are indicative of better health-related quality of life (high = good).

For endpoints evaluating a single question, the raw score is used in the analysis. The activity and revised activity domain scores will be summarized by treatment group.

Scale	Sum Item Values	Lowest and Highest Possible Raw Scores	Possible Raw Score Range
Concern	9+15+22+28+32	5, 25	20
Activities	10+11+13+19+20+27+29	7, 35	28
Revised activities	11+13+19+20+27	5,25	20
Energy/mood	12+17+23+24+25+31+35	7, 35	28
Control	14+16+26+30+34	5, 25	20
Self-conscious	18+21+33	3, 15	12
Sexual function	36+37	2, 10	8
HRQL TOTAL	Sum of 6 Subscale Scores[a]	29, 145	116

Abbreviations: HRQL, health-related quality of life.

[a] HRQL Total includes following scales: concern, activities, energy/mood, control, self-conscious, and sexual function.

Formula for Transformation of Raw Scores of Other Scale Scores ONLY:

$$\text{Transformed Score} = [(\text{Highest possible score} - \text{Actual raw score})/(\text{Possible raw score range})] * 100$$

For revised activities, the proportion of patients who are responders (defined as meeting a meaningful change from Baseline in the revised activity score) at Week 24 will be analyzed similarly to that for the change in BPD Scale score between the two treatment arms (relugolix + E2/NETA and placebo). The proposed responder threshold is a 20-point increase. Details of the determination of the meaningful change in the Revised Activities Scale score can be found in Appendix 5.

Missing Items

For any scale analyses, if < 50% of the scale items are missing, the scale should be retained using the mean scale score of the items present. If ≥ 50% of the items are missing, no scale score should be calculated; the subscale score will be considered missing.

7.4.4.3. Patient Global Assessment

The PGA for function and symptoms will be evaluated using a 5-point response scale (eg, absent, mild, moderate, severe, and very severe). To calculate change from Baseline to Week 24, the following numerical scores will be assigned to each response level:

Response Scale (Function)	Response Scale (Symptoms)	Numerical Score
No limitation at all	Not severe	1
Mild limitation	Mildly severe	2
Moderate limitation	Moderately severe	3
Quite a bit of limitation	Very severe	4
Extreme limitation	Extremely severe	5

For each item, the count and proportion of improvement by level or at least one level will be tabulated by treatment group and by visit. The denominator for the proportion will be based on the number of patients who provided non-missing responses to the items.

7.4.4.4. Menorrhagia Impact Questionnaire

The Menorrhagia Impact Questionnaire items 3 and 4 will be evaluated using the 5-point response scales (Not at all, Slightly, Moderately, Quite a bit, and Extremely) to assess level of improvement from Baseline to Week 24.

For each item, the count and proportion of improvement by level will be tabulated by treatment group and by visit. The denominator for the proportion will be based on the number of patients who provided non-missing responses to the items.

7.5. Exploratory Efficacy Endpoints

The following exploratory endpoints will be assessed for both comparisons the relugolix + E2/NETA group with the placebo group and the relugolix + delayed E2/NETA group with the placebo group:

- Change from Baseline to Week 24 in the EQ-5D-5L Scale score
- Change from Baseline to Week 24 in EQ-5D-5L visual analogue score.

7.5.1. Exploratory Efficacy Analyses

Analysis methods previously described for primary and secondary efficacy endpoint analyses will be used for the analysis of these endpoints.

8. PHARMACOKINETIC AND PHARMACODYNAMIC ANALYSES

Plasma relugolix, plasma NET, and serum E2 trough concentrations will be listed and summarized by study, treatment group (Group A, B, or C), and visit.

Serum pharmacodynamic data (LH, FSH, E2, and progesterone) will be listed and summarized using descriptive statistics (including raw and change from Baseline) by study, treatment group (Group A, B, or C), and visit.

For pharmacodynamic assessment, the number and percentage of patients with individual E2 concentration values < 10 pg/mL, 10 to < 20 pg/mL, 20 to < 50 pg/mL, and ≥ 50 pg/mL and individual progesterone concentration values < 1 ng/mL, 1 to 5 ng/mL, and ≥ 5 ng/mL will be summarized by treatment group (Group A, B, or C) and visit.

Scatter plots with LOESS smoothing lines for MVT-601-3001 and MVT-601-3002 separately will be used to examine the relationship between mean plasma relugolix trough concentration at the given time point (collected between 18 and 30 hours after the previous dose) and the following pharmacodynamic concentrations:

- Week 12 serum LH, FSH, E2, and progesterone (separately for Groups A and B);
- Week 24 serum LH, FSH, E2, and progesterone (separately for Groups A and B, and Groups A and B combined).

In addition, the PK data from this study will be combined with PK data from other studies to define a population PK model, which will be reported separately. Exposure-response analyses of the primary efficacy endpoint and safety will be conducted to assess the effect of relugolix exposure on outcomes. The analysis plan for population PK and exposure-response analyses will be specified in a separate document.

9. SAFETY ANALYSES

Unless otherwise specified, safety analyses will be conducted using the safety population according to the treatment received by the patients.

9.1. Adverse Events

Adverse events will be collected from the time of the first dose of study drug through the safety follow up visit approximately 30 days after the last dose of study drug (the end of treatment period), or the date of initiation of another investigational agent or hormonal therapy or surgical intervention or entering extension study, whichever occurs first. Serious adverse events reported to the investigator after the safety reporting period should be reported to the sponsor if the investigator assesses the event as related to study drug.

The severity of all treatment-emergent adverse events will be evaluated by the investigator based on the National Cancer Institute's Common Terminology Criteria for Adverse Events (CTCAE) and will be coded to preferred term and system organ class using MedDRA 22.0 or higher.

A treatment-emergent adverse event is defined as any adverse event that occurs after administration of the first dose of study drug.

Adverse event summaries will be based on treatment-emergent adverse events, unless otherwise specified. Adverse events occurring prior to administration of any study drug will be listed and flagged in by-patient listings.

The following tabular summaries that include the number and percentage of patients will be provided:

- Overview of adverse events;
- All adverse events;
  - By SOC and PT;
  - By decreasing frequency of PT;
  - By SOC, PT, and maximum severity;
  - Study drug-related per investigator by SOC and PT;
  - By time to onset, SOC and PT;
- Grade 3 or above adverse events;
  - By SOC and PT;
  - By decreasing frequency of PT;
  - Study drug-related per investigator by SOC and PT;
- Grade 2 or above adverse events;
  - By SOC and PT;
  - By decreasing frequency of PT;
  - By SOC, PT, and maximum severity;

  - Study drug-related per investigator by SOC and PT;
- Adverse events leading to study drug withdrawal;
  - By SOC and PT;
  - By decreasing frequency of PT;
- Adverse events leading to dose interruption;
  - By SOC and PT;
  - By decreasing frequency of PT;
- Adverse events resulting in fatal outcome;
  - By decreasing frequency of PT;
- Serious adverse events;
  - By SOC and PT;
  - By decreasing frequency of PT;
  - By SOC, PT, and maximum severity;
  - By SOC, PT, and relationship to study drug;
- Adverse events of clinical interest (ALT or AST ≥ 3 x ULN);
  - By SOC, PT, and maximum severity;
  - By decreasing frequency of PT.

Additionally, adverse event categories defined in Table 12 will be summarized by decreasing frequency of PT.

9.1.1. Relationship to Study Drug

Adverse events will be classified as "related" to study treatment if the relationship was rated by the investigator as possibly related or probably related. Adverse events related to any study drug (relugolix or placebo and E2/NETA or placebo) will be considered as related to study drug.

9.1.2. Severity of Adverse Event

Grade 2 or above adverse events will be summarized by SOC, PT, and/or maximum severity, relationship to study treatment.

9.1.3. Serious Adverse Event

Serious adverse events will be summarized by SOC, PT, and/or maximum severity, relationship to study treatment.

The data handling conventions for and the definition of a serious adverse event are discussed in this section. All deaths during the study, including the post treatment follow-up period, and deaths that resulted from a process that began during the study, should be included in the

analysis. For more details, deaths occurring during the following time periods or under the following conditions should be considered:

- Deaths occurring during participation in any study, or during any other period of drug exposure
- Deaths occurring after a patient leaves a study, or otherwise discontinues study drug, whether or not the patient completes the study to the nominal endpoint, if the death:
  - Is the result of a process initiated during the study or other drug exposure, regardless of when it actually occurs; or
  - Occurs within a time period that might reflect drug toxicity for a patient leaving a study or otherwise discontinuing drug. For drugs with prompt action and relatively short elimination half-lives, 4 weeks is a reasonable time period. For drugs with particularly long elimination half-lives or drug classes with recognized potential to cause late occurring effects, deaths occurring at longer times after drug discontinuation should be evaluated.

9.1.4. Adverse Event Leading to Withdrawal of Study Drug

Adverse events leading to withdrawal of study drug are those adverse events collected from the adverse event CRF pages with “drug withdrawn” as the action taken with study drug.

Adverse events with “drug withdrawn” as action taken due to any one of the components of study drug will be considered as leading to withdrawal of study drug.

9.1.5. Adverse Events Leading to Dose Interruption

Adverse events leading to dose interruption are those adverse events collected from the adverse event CRF pages with “drug interrupted” as their action taken with study drug.

Adverse events with “drug interrupted” as action taken due to any one of the components of study drug will be considered as leading to dose interruption.

9.1.6. Adverse Events Resulting to Fatal Outcome

Adverse events resulting in a fatal outcome are those adverse events collected from the adverse event pages with “fatal” as their outcome.

The fatal events, if any, will be provided in a by-subject listing.

9.1.7. Adverse Event Categories

In addition, adverse event categories defined in Table 12 will be summarized by decreasing frequency of PT under each safety population. Incidence of vasomotor symptoms by 12 weeks will be compared between relugolix Group A and relugolix Group B. Comparative statistics (such as p-values, 95% CIs, risk ratio) will be provided. Vasomotor symptoms throughout the studies will be summarized by SOC, PT, and maximum severity.

Table 12: Constitution of Adverse Event Categories

Category	Search Criteria
Bone health events	Osteoporosis/Osteopenia SMQ (broad) Fracture (custom SMQ): All preferred terms including the term "fracture," excluding "Tooth fracture" and "Fracture of penis"
Hepatic disorders	Drug-related hepatic disorders – comprehensive SMQ (narrow)
Metabolic disorders	Dyslipidemia SMQ (broad) Hyperglycemia/new onset diabetes mellitus SMQ (narrow)
Vasomotor symptoms	The following 5 preferred terms will be included: Hyperhidrosis; Feeling hot; Hot flush; Night sweats; Flushing
Mood disorders	MedDRA Depression and Suicide/Self-Injury SMQ (broad)

Abbreviations: HLT, High-Level Term; MedDRA, Medical Dictionary for Regulatory Activities; SMQ, Standardised MedDRA Query.

9.2. Laboratory Data

Laboratory parameters, including chemistry and hematology panels, specified as per protocol, and collected from the central laboratory will be tabulated and presented in by-patient listings. Urinalysis and hepatitis virus serological test results will be provided in by-patient listing only.

The National Cancer Institute CTCAE Grading Scale with numeric component will be used to categorize toxicity grade for laboratory parameters (CTCAE v5.0, dated 17 Nov 2017). Parameters that have criteria available for both low and high values (eg, hypercalcemia for a high value of calcium and hypocalcemia for a low value of calcium) will be summarized for both criteria (low and high). Patients will only be counted once for each criterion. The same patient can be counted for both criteria if she has laboratory values meeting each criterion. Shift tables will be provided for each gradable parameter to summarize Baseline toxicity grade versus worst post-Baseline toxicity grade. For laboratory parameters that are not gradable by the CTCAE, a shift table based upon the normal range (low, normal, and high) will be provided for each parameter to summarize the Baseline versus worst post-Baseline results.

Boxplots of laboratory values over time will be plotted for key laboratory parameters. These laboratory parameters include, but are not limited to, hematology (hemoglobin, platelets,

leukocytes, neutrophils), creatinine, glomerular filtration rate, and hepatic function panel (alanine aminotransferase [ALT], aspartate aminotransferase [AST], alkaline phosphatase [ALP], and total bilirubin).

The change from Baseline to each post-Baseline study visit will be presented by treatment group for each laboratory test in both tables and figures.

The number and proportion of patients with liver test elevations will be presented by treatment group. Liver test elevations are assessed by using post-Baseline results for ALT, AST, ALP, and total bilirubin based on the definitions presented in Table 13.

Table 13: Categories of Liver Test Elevations

Laboratory Test	Category
ALT or AST	ALT or AST > ULN - < 3xULN ALT or AST ≥ 3x to < 5x ULN ALT or AST ≥ 5x to < 8x ULN ALT or AST ≥ 8x to < 10x ULN ALT or AST ≥ 10 to < 20x ULN ALT or AST ≥ 20x ULN
Total bilirubin	Total bilirubin > 2 × ULN
ALT or AST and total bilirubin	ALT or AST ≥ 3 × ULN + total bilirubin > 2 × ULN
ALT or AST, total bilirubin, and ALP	ALT or AST ≥ 3 x ULN + total bilirubin > 2 × ULN + ALP < 2 × ULN

Abbreviations: ALP, alkaline phosphatase; ALT, alanine aminotransferase; AST, aspartate aminotransferase; ULN, upper limit of normal.

The number and percentage of patients with concurrent (defined as measurements on the same day) ALT or AST ≥ 3 times ULN and total bilirubin > 2 times ULN will also be presented.

9.3. Other Safety Analyses

9.3.1. Electrocardiograms

ECG interval results and changes from Baseline will be summarized descriptively for each scheduled visit in both tables and figures using data provided by and read by central reading.

A categorical analysis of corrected QT interval using Fridericia's calculation (QTcF) intervals will also be performed for each scheduled visit and for the maximum post-Baseline value. The number and percentage of patients in each QTcF interval category (< 450 msec, 450 to 480 msec, 481 to 500 msec, and > 500 msec) will be summarized. Categories of changes from Baseline (≥ 30 msec and ≥ 60 msec) will be summarized as well.

ECG intervals will be presented in by-patient listing. Overall ECG assessments performed by local reading will also be listed.

9.3.2. Visual Acuity

Visual Acuity Score at Baseline and at each scheduled post-Baseline assessment time point will be presented in a by-patient listing.

9.3.3. Vital Signs

Blood pressure (systolic and diastolic), heart rate, and BMI will be summarized at Baseline and each subsequent scheduled assessment by treatment group. Change from Baseline will be calculated and presented for each parameter at all scheduled post-Baseline assessment time points in both tables and figures. All vital sign data will also be provided in by-patient listings.

Potentially clinically significant abnormalities in vital signs are defined in Table 14, and they will be summarized by using post-Baseline values that meet the defined criteria. Potentially clinically significant abnormalities will also be flagged in by-patient listings.

Table 14: Categories of Potentially Clinically Significant Abnormalities in Vital Signs

Parameter	Category
Systolic blood pressure	≥ 140 mmHg ≥ 180 mmHg
	≤ 90 mmHg
	Increase of ≥ 20 mmHg from Baseline Decrease of ≥ 20 mmHg from Baseline
Diastolic blood pressure	≥ 90 mmHg ≥ 105 mmHg
	≤ 50 mmHg
	Increase of ≥ 15 mmHg from Baseline Decrease of ≥ 15 mmHg from Baseline
Heart rate	≥ 120 bpm
	< 45 bpm
	Increase of ≥ 15 bpm from Baseline Decrease of ≥ 15 bpm from Baseline

Abbreviations: bpm, beats per minute; mmHg, millimeters of mercury.

9.3.4. Endometrial Biopsy

Primary diagnosis of endometrial biopsy assessment will be summarized at Baseline and at scheduled assessment by treatment group. All endometrial biopsy data will also be provided in a by-patient listing.

Primary diagnosis from pathologist evaluation will be categorized by medical monitor's review in Table 15 and will be summarized using frequencies and percentages, summarized for each treatment group. All endometrial biopsy data will also be provided in by-patient listings.

Table 15: Categories of Primary Diagnosis in Endometrial Biopsies

Normal-Proliferative	• Weakly proliferative • Proliferative • Disordered proliferative
Normal-Secretory/Menstrual/Mixed	• Secretory • Menstrual • Progestational/Decidulized/Mixed
Normal-Atrophic or Minimally Stimulated	• Atrophic • Indeterminate/Inactive
Hyperplasia	• Simple hyperplasia without atypia • Simple hyperplasia with atypia • Complex hyperplasia without atypia • Complex hyperplasia with atypia
Carcinoma	—
Inadequate	—
Missing	—
Additional Diagnosis (Other reported finding)	• Reactive/Inflammatory • Polyp • Metaplasia • Glandular and/or Stromal Breakdown

9.3.5. Bone Mineral Density

Corrected BMD data will be used for analysis as determined by the central radiology laboratory in the 3 prespecified anatomical locations: lumbar spine (L1–L4), total hip, and femoral neck.

BMD at Baseline, Week 12 and Week 24 visits will be summarized descriptively by treatment group and each anatomical location. Percentage changes from Baseline along with 95% CIs of mean percentage changes will be also summarized by treatment group and anatomical location. Mean percentage change from Baseline with its corresponding 95% CI will be plotted by visit, treatment group, and anatomical location.

To support the inclusion of E2/NETA in the treatment regimen, the safety endpoint of mean percent change from Baseline in BMD at the lumbar spine at Week 12 will be analyzed using pooled data from the two replicate studies (MVT-601-3001 and MVT-601-3002) with a formal

comparison of the relugolix + E2/NETA group (Group A) versus the relugolix + delayed E2/NETA group (Group B) (details in the Integrated Summary of Safety Statistical Analysis Plan).

In addition, the difference of percentage change from Baseline between treatment groups (relugolix + E2/NETA group versus the relugolix + delayed E2/NETA group at 12 weeks, relugolix + E2/NETA versus placebo group at 12 and 24 weeks, and relugolix + delayed E2/NETA group versus placebo group at 12 weeks) will be summarized at each visit by anatomical location along with the corresponding 95% CIs.

To account for participants whose BMD assessment may have been obtained outside of the protocol-specified window (Week 12 ± 3 weeks, Week 24 ± 3 and 4 weeks), a sensitivity analysis by visit will be conducted that includes all women who underwent DXA at both time points, regardless of whether the image was procured during the prespecified time window.

A mixed-effects model with repeated measures will be used to describe treatment effect on BMD at 12 and 24 weeks. The model will have treatment group, age at Baseline, visit, Baseline BMD value, stratification factors (geographic region and menstrual blood loss volume), race (African American versus Other), and BMI at Baseline as fixed effects using an unstructured variance-covariance matrix. Least square means on each anatomical location will be presented and plotted at each visit with associated 95% CIs. Categorical representation of percentage change from Baseline to 12 and 24 weeks of treatment will be presented by the number and proportion of patients who had BMD declines of ≤ 2%, >2% to 3%, > 3% to 5%, > 5% to 8%, and > 8% by treatment group and anatomical location. The 95% CIs will be provided for the respective proportions.

Categorical changes from Baseline in overall BMD (defined as lumbar spine and total hip) also will be assessed at 12 and 24 weeks. Femoral neck evaluates a smaller area of bone mass than the total hip and is prone to lower precision in the measurement (ISCD Official Positions, 2015; Leslie, 2007). Since femoral neck BMD may be associated with discordant readings compared with the total hip or lumbar spine due to technical considerations, it will not add meaningful interpretation of overall BMD changes in response to treatment.

Z-scores will be summarized by treatment group, visit, and anatomical location with descriptive statistics including 95% CIs, and the number and percentage of patients with a Z-score < -2.0 will be presented by treatment group, visit, and anatomical location.

BMD percentage changes from Baseline will also be summarized by intrinsic factors (eg, age, race, body mass index) and extrinsic factors (eg, geographic region).

9.3.6. Bleeding Pattern

Bleeding patterns will be summarized at Week 24/EOT by treatment group. Three bleeding patterns will be considered: amenorrhea (see Section 7.4.3), cyclic bleeding, and irregular bleeding. Patients with the cyclic bleeding pattern are those who do not meet the definition of amenorrhea and do meet the following conditions:

- 3 to ≤ 12 days of menstruation duration per eDiary at Week 24/EOT window (see Section 7.3.3)

- No more than 2 days of gap of no bleeding (per eDiary) within the menstruation duration.

Patients with the irregular bleeding pattern are those who do not meet the definitions of cyclic bleeding or amenorrhea. The number (and percent) of patients and mean number of bleeding days will be provided by treatment group for each bleeding pattern.

For patients with cyclic or irregular bleeding pattern, the number (and percent) of patients with observed MBL volume falling into the following bleeding intensity groups will be provided:

- **Spotting/negligible bleeding:** MBL volume < 5 mL
- **Light:** MBL volume 10 - 50 mL
- **Moderate:** MBL volume >50 to ≤80 mL
- **Heavy:** MBL volume > 80 mL

For each bleeding intensity category, the mean number of bleeding days will be summarized.

10. REFERENCES

2015 International Society for Clinical Densitometry (ISCD) Official Positions – Adult (https://www.iscd.org/official-positions/2015-iscd-official-positions-adult/; accessed 30 Apr 2019).

Bachowski G, Borge D, Brunker P, Eder A, Fialkow L, Friday J, et al. A Compendium of Transfusion Practice Guidelines. American Red Cross; 2017 p. 81. Report No.: 3rd Edition.

Leslie WD, Lix LM, Tsang JF, Caetano PA. Single-site vs Multisite Bone Density Measurement for Fracture Prediction. Arch Intern Med 2007; 167 (15): 1641-7

Man L, Tahhan HR. Body surface area: a predictor of response to red blood cell transfusion. JBM. 2016 Sep;Volume 7:199–204.

Ratitch B, Lipkovich I, O'Kelly M. Combining analysis results from multiply imputed categorical data. PharmaSUG 2013. Paper SP03.

Rubin D. The calculation of posterior distributions by data augmentation: comment: a noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: the SIR algorithm. J Am Stat Assoc. 1987 Jun; 82(398), 543-546.

Stewart EA, Owens C, Duan WR, Gao J, Chwalisz K, Simon JA. Elagolix alone and with add-back decreases heavy menstrual bleeding in women with uterine fibroids. Poster presented to American College of Obstetricians and Gynecologists Annual Clinical and Scientific Meeting 2017, San Diego, CA, May 6-9, 2017.

Sullivan TR, White IR, Salter AB, Ryan P, Lee KJ. Should multiple imputation be the method of choice for handling missing data in randomized trials? Stat Methods Med Res. 2018 Sep; 27(9):2610-2626.

von Hippel PT. How many imputations do you need? A two-stage calculation using a quadratic rule. Sociol Methods Res. 2018 Jan.

APPENDICES

APPENDIX 1. SUMMARY OF SECONDARY ENDPOINT ANALYSES

Secondary Endpoints	Analysis Population	Statistical Method/Test	Declare Statistical Significance [a]	Time Points of Summary	Summary Statistics
Key Secondary Efficacy Endpoints with Alpha Protection					
Proportion of women who achieve amenorrhea over the last 35 days of treatment	mITT	Cochran-Mantel-Haenszel test	P < 0.05	Week 24	Frequency and percentages
% change from Baseline to Week 24 in MBL volume	mITT	Mixed-effects model	P < 0.05	Week 24	LS means for % change
Proportion of women with a hemoglobin ≤10.5 g/dL at Baseline who achieve an increase of > 2 g/dL from Baseline at Week 24	Subset of mITT	Cochran-Mantel-Haenszel test	P < 0.05	Week 24	Frequency and percentages
Change from Baseline to Week 24 in the UFS-QoL Bleeding and Pelvic Discomfort Scale score, a sub-scale of the UFS-QoL Symptom Severity Scale	mITT	Mixed-effects model	P < 0.05	Week 24	LS means for change
Proportion of patients with a maximum NRS score ≤ 1 during the 35 days before the last dose of study drug in the subset of women with a maximum NRS score ≥4 for pain associated with uterine fibroids during the 35 days prior to randomization	Subset of mITT	Cochran-Mantel-Haenszel test	P < 0.05	Week 24/EOT	Frequency and percentages
% change from Baseline to Week 24 in uterine fibroid volume	mITT	ANCOVA model	P < 0.05	Week 24	LS means for % change
% change from Baseline to Week 24 in uterine volume	mITT	ANCOVA model	P < 0.05	Week 24	LS means for % change
Other Secondary Efficacy Endpoints					
Time to achieve MBL volume of < 80 mL AND at least a 50% reduction from Baseline MBL volume as measured by the alkaline hematin method	mITT	Log-rank test/KM method	P < 0.05	Monthly	KM estimates at Week 12 and 24, KM plots, median time to response
Time to achieve amenorrhea	mITT	Log-rank test/KM method	P < 0.05	Monthly	KM plots, median time to response

Secondary Endpoints	Analysis Population	Statistical Method/Test	Declare Statistical Significance [a]	Time Points of Summary	Summary Statistics
Time to sustained amenorrhea	mITT	Log-rank test/KM method	P < 0.05	Monthly	KM plots, median time to response
Proportion of women in the relugolix Group A versus the placebo Group C who achieve an MBL volume of < 80 mL AND at least a 50% reduction from Baseline MBL volume at Week 4, Week 12, Week 16, and Week 20	mITT	No comparison		*at Week 4, Week 12, Week 16, and Week 20*	Descriptive
Sustained amenorrhea rate by visit	mITT	Cochran-Mantel-Haenszel test	P < 0.05	Monthly	Frequency and percentages
Proportion of women with a hemoglobin below the lower limit of normal at Baseline who achieve an increase of ≥ 1 g/dL from Baseline at Week 24	mITT	Cochran-Mantel-Haenszel test	P < 0.05	Monthly	Frequency and percentages
Change (absolute and %) from Baseline to Week 24 in hemoglobin for women with a hemoglobin ≤ 10.5g/dL at Baseline	mITT	Mixed-effects model	P < 0.05	Monthly	LS means for % change
Proportion of women who achieve a maximum Numerical Rating Scale score for uterine fibroid-associated pain over the last 35 days of treatment that is at least a 30% reduction from Baseline in the subset of women with a maximum pain score ≥ 4 during the 35 days prior to randomization	Subset of mITT	Cochran-Mantel-Haenszel test	P < 0.05,	Week 12, 24	Frequency and percentages
Mean maximum NRS scores over time	Subset of mITT	Descriptive		Monthly	Means
Proportion of responders who had meaningful reduction of >20 points from Baseline to Week 24 in UFS-QOL Bleeding and Pelvic Discomfort Scale (Q1, Q2 and Q5)	mITT	Cochran-Mantel-Haenszel test	P < 0.05,	Week 12, 24	Frequency and percentages
Proportion of responders who had meaningful increase of > 20 points from Baseline to Week 24 in UFS-QOL revised activities	mITT	Cochran-Mantel-Haenszel test	P < 0.05,	Week 12, 24	Frequency and percentages

Secondary Endpoints	Analysis Population	Statistical Method/Test	Declare Statistical Significance [a]	Time Points of Summary	Summary Statistics
Change from Baseline to Week 24 in impact of uterine fibroids based on the UFS-QOL revised activities domain	mITT	Mixed-effects model	P < 0.05	Week 12, 24	LS means for change
Change from Baseline to Week 24 in impact of uterine fibroids based on the UFS-QOL activities domain	mITT	Mixed-effects model	P < 0.05	Week 12, 24	LS means for change
Change from Baseline to Week 24 in the interference of uterine fibroids with physical activities based on UFS-QOL Q11	mITT	Mixed-effects model	P < 0.05	Week 12, 24	LS means for change
Change from Baseline to Week 24 in the interference of uterine fibroids with social activities based on UFS-QOL Q20	mITT	Mixed-effects model	P < 0.05	Week 12, 24	LS means for change
Change from Baseline to Week 24 in embarrassment caused by uterine fibroids based on UFS-QOL Q29	mITT	Mixed-effects model	P < 0.05	Week 12, 24	LS means for change
Change from Baseline to Week 24 in the UFS-QoL Symptom Severity Scale score	mITT	Mixed-effects model	P < 0.05	Week 12, 24	LS means for change
Change from Baseline to Week 24 in the UFS-HRQL total score	mITT	Mixed-effects model	P < 0.05	Week 12, 24	LS means for change
Change in PGA for uterine fibroid related function from Baseline to Week 24	mITT	Mixed-effects model	P <0.05	Monthly	LS means for absolute and change
Change in PGA for uterine fibroid symptoms from Baseline to Week 24	mITT	Mixed-effects model	P < 0.05	Monthly	LS means for absolute and change
Proportion of patients achieving improvement in PGA for uterine fibroid symptoms from Baseline to Week 24	mITT	Cochran-Mantel-Haenszel test	P < 0.05	Monthly	Frequency and percentages
Proportion of patients achieving improvement in PGA for uterine fibroid related function from Baseline to Week 24	mITT	Cochran-Mantel-Haenszel test	P < 0.05	Monthly	Frequency and percentages
Safety Related Endpoints					
% Change from Baseline to Week 12 in BMD (pooled data)	Safety population	Mixed-effects model Relugolix Group A vs B	P < 0.05,	Week 12	LS means Diff (95%CI)

Secondary Endpoints	Analysis Population	Statistical Method/Test	Declare Statistical Significance [a]	Time Points of Summary	Summary Statistics
% Change from Baseline in BMD	Safety population	Mixed-effects model Relugolix Group A vs Placebo at 12/24 weeks; Relugolix Group B vs Placebo at 12 weeks		Week 12, 24	LS means Diff (95%CI)
Exploratory Secondary Efficacy Endpoints					
Change from Baseline to Week 24 in the Menorrhagia Impact Questionnaire Score for physical activities	mITT	Descriptive		Monthly	Frequency and percentages
Change from Baseline to Week 24 in the Menorrhagia Impact Questionnaire Score for social and leisure activities	mITT	Descriptive		Monthly	Frequency and percentages

Abbreviations: KM, Kaplan-Meier; LS, least squares; mITT, modified intent-to-treat; NRS, Numerical Rating Scale; Q, question; UFS-HRQL, Uterine Fibroid Scale – Health-related Quality of Life.
[a] P-values are two-sided.

APPENDIX 2. DERIVATION AND PSYCHOMETRIC EVALUATION OF A UFS-QOL BLEEDING AND PELVIC DISCOMFORT SCALE

The BPD Scale was derived from the Symptom Severity Scale of the Uterine Fibroid Symptom and Health-Related Quality of Life (UFS-QoL). The BPD Scale consists of three items proximal to uterine fibroids that are experienced by most patients, (ie, heavy bleeding during the menstrual period [Question 1], passing blood clots during the menstrual period [Question 2], and feeling tightness or pressure in the pelvic area [Question 5]).

The aim of this appendix is to describe the derivation and psychometric testing process of the BPD Scale. Results of the analyses in this appendix are summarized in Appendix 3 and will be included in the Patient-Reported Outcomes dossier to be submitted at the time of filing for the uterine fibroids registration program.

Exploratory factor analysis and subsequent confirmatory factor analysis were conducted to assess and confirm the factor structure of the Symptom Severity Scale of the UFS-QoL, using data from a phase 2 study of relugolix in uterine fibroids (TAK-385/CCT-001), as well as pooled, blinded data from one-third of patients in the phase 3 studies (MVT-601-3001 and MVT-601-3002). Respective analyses are described in Section 2.1. Based on the results, the factor(s) reflecting symptoms proximal to uterine fibroids and experienced by most patients with uterine fibroids were selected for further psychometric testing.

The psychometric properties of the new scale were assessed using the same pooled, blinded data from the two phase 3 studies of relugolix in uterine fibroids (MVT-601-3001 and MVT-601-3002). These analyses are described in Section 2.2. The blinded data consists of the first third of patients (approximately n = 260) enrolled into the two pivotal studies who have completed the patient global assessment (PGA) for symptoms and the UFS-QoL at Baseline and at Week 24. Of note, for the analyses specified in Section 2.2, only data at Baseline and Week 12 were used; the Week 24 data was used in the responder analyses described in Appendix 3.

2.1. Development of the Bleeding and Pelvic Discomfort Scale Using Phase 2 and Phase 3 Data

From a review of the eight items in the Symptom Severity Scale of the UFS-QoL, it was apparent that the scale consists of different constructs/dimensions. Therefore, the factor structure of the Symptom Severity Scale was assessed, initially using data from the phase 2 study TAK-385/CCT-001 (n = 216).

Of note, in the TAK-385/CCT-001 phase 2 study, the UFS-QoL with a one-month recall period was applied, whereas the UFS-QoL with a three-month recall period is used in the phase 3 studies (MVT-601-3001 and MVT-601-3002). Therefore, confirmatory factor analysis and final psychometric testing of the chosen factor was conducted using blinded phase 3 data (see Section 2.2).

Exploratory Factor Analysis

The exploratory factor analysis was done on phase 2 data to identify the underlying constructs by the most parsimonious factor structure of the eight items in the Symptom Severity Scale. Identification of the number of factors was based on the following criteria:

- Items with primary factor loading > 0.4;
- Factors with large eigenvalues considered as common factors using Kaiser criterion (Kaiser, 1960).

A scree plot was used as a supplemental tool to decide on the number of factors in the final model.

Confirmatory Factor Analysis

Once the number of factors was identified, a confirmatory factor analysis was conducted using blinded, pooled phase 3 data to confirm the factor structure. Only patients who completed the Baseline and Week 24 PGA for symptoms and UFS-QoL assessments were included in this analysis. Model fit was assessed based on the following:

- The goodness of fit as measured by χ^2 and Goodness of Fit Index; a Goodness of Fit Index > 0.9 is considered acceptable;
- The Comparative Fit Index was used to determine the acceptability of the model fit of the discrepancy function adjusted sample size; a Comparative Fit Index > 0.9 (Hu, 1995) was considered an acceptable fit;
- The root mean square error of approximation was used to determine the acceptability of model fit of the square root of the discrepancy between the sample covariance matrix and the model covariance matrix; the root mean square error of approximation had to be < 0.06 (Browne, 1993) to be considered an acceptable fit;
- P-value > 0.05.

Once the final factor structure was identified, the factor reflecting items proximal to uterine fibroids and experienced by almost all patients with uterine fibroids were selected for further evaluation. Of note, this was the BPD Scale.

2.2. Psychometric Analyses Based on Phase 3 Data

The same pooled, blinded data from the first third of patients enrolled in either of the two phase 3 studies (MVT-601-3001 or MVT-601-3002) was used for the psychometric analyses of the BPD Scale. The objective was to psychometrically evaluate the new scale in terms of item performance, reliability, validity, and ability to detect change. Of note, for the analyses specified in this section, only data at Baseline and Week 12 were used. The following analyses were performed:

Item Level Analysis Assessing Ceiling and Floor Effects:

- A descriptive summary of the eight items in the UFS-QoL Symptom Severity Scale at Baseline was provided to examine item distributions and ceiling/floor effects. Low ceiling effects (< 20%) and higher floor effects (> 20%) were expected at Baseline

due to symptom severity of patients with uterine fibroids enrolled in the phase 3 studies.

Internal Consistency:

Internal consistency reliability was assessed for the BPD Scale at Baseline and Week 12 by calculating Cronbach's alpha. Generally, a Cronbach's alpha coefficient (α) $\geq$ 0.7 indicates an acceptable level of internal consistency.

Item Performance:

- Intercorrelation of items that contribute to the BPD Scale by means of item-total correlation was determined.
- Item discrimination index was assessed for each item based on 1) the BPD Scale scores at single time points, and 2) the change from Baseline to Week 12 in the BPD Scale score to determine the degree to which individual items were able to discriminate between less and more severe patients (Cappelleri, 2014).

Known-Groups Validity:

- Known-groups validity was assessed based on groups defined by Baseline PGA for symptoms severity (five levels). Descriptive statistics of the BPD Scale will be provided for each severity level.

Ability to Detect Change:

Evidence that the new scale can identify differences in scores over time in individuals or groups who have changed with respect to the measurement concept will be investigated by providing the following descriptive statistics:

- Within person change from Baseline to Week 12 in each item on the BPD Scale
- Standardized effect size statistic (SES) for change from Baseline to Week 12 in each item scale. The ability to detect change will be judged based on Cohen's recommendations: small change (SES = 0.20), moderate change (SES = 0.50), and large change (SES = 0.80).

2.3. References

Browne MW, Cudeck R. Alternative ways of assessing model fit. In: Bollen KA, Long JS (eds), Testing structural equation models (Vol. 154, pp. 136-162). 1993. Newbury Park, CA: Sage Focus Editions.

Cappelleri JC, Bushmakin AG. Interpretation of patient-reported outcomes. Stat Methods Med Res. 2014;23(5), 460–483.

Hu LT, Bentler PM. Evaluating model fit. In: Hoyle RH (ed), Structural equation modeling: concepts, issues, and applications (pp. 76-99). 1995. Thousand Oaks, CA, US: Sage Publications, Inc.

Kaiser HF. The application of electronic computers to factor analysis. Educ Psychol Meas. 1960;20:141-151.

APPENDIX 3. DERIVATION AND VALIDATION OF THE UFS-QOL BLEEDING AND PELVIC DISCOMFORT SCALE

Results described in this appendix are based on the analyses described in Appendix 2.

3.1. Development of the Bleeding and Pelvic Discomfort Scale Using Exploratory and Confirmatory Factor Analysis

Exploratory factor analysis was conducted on data from the phase 2 study TAK-385/CCT-001 study (n = 216) and the arising factor structure was assessed in a confirmatory factor analysis using data from the phase 3 studies MVT-601-3001 and MVT-601-3002.

3.1.1. Exploratory Factor Analysis Using Phase 2 Data

Exploratory factor analysis results revealed a two-factor solution based on the Kaiser criterion (eigenvalues > 1) and factor loading > 0.40 criteria specified in the analysis plan (see Appendix 2). Factor 1 and Factor 2 had eigenvalues of 3.394 and 1.196, respectively (Table 3.1-1). Three items were found to load adequately onto Factor 1 with loadings greater than 0.40: Item 1 (Heavy Bleeding during Your Period), Item 2 (Passing Blood Clots during Your Period), and Item 5 (Feeling Tightness or Pressure in Pelvis; see Table 3.1-2). Two items loaded onto Factor 2 with loadings larger than the prespecified level: Item 6 (Frequent Urination in Daytime) and Item 7 (Frequent Nighttime Urination). Item 8 (Feeling Fatigued) showed a loading value on Factor 1 just below the prespecified threshold (0.399) and showed evidence of cross-loading with the Factor 2 (0.288). An additional factor with a moderate eigenvalue (0.62) was considered based the scree plot (Figure 3.1-1) and factor loadings of its associated items (Item 3: Fluctuation in Duration of Menstruation, 0.416; Item 4: Fluctuation in Length of Monthly Cycle, 0.995; Table 3.1-2).

Overall the results show support for a seven-item three-factor model. Due to multi-factor loading, Item 8 (Feeling Fatigued) remains a single-item symptom and is not scored as part of any factor.

Figure 3.1-1: Scree Plot and Variance Explained for UFS-QoL Symptom Severity Scale Factors in TAK-385/CCT-001

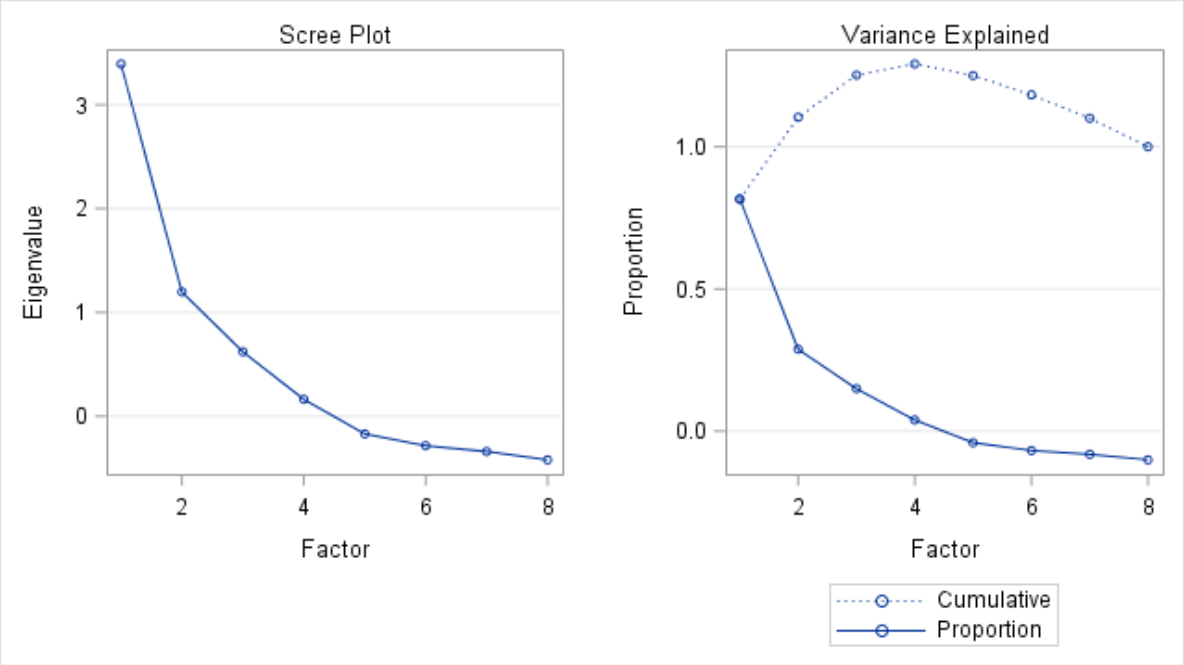


Table 3.1-1: Exploratory Factor Analysis for the UFS-QoL Symptom Severity Scale in TAK-385/CCT-001

Item	Eigenvalue	Difference	Proportion	Cumulative
1	3.394	2.198	0.816	0.816
2	1.196	0.576	0.288	1.104
3	0.620	0.458	0.149	1.253
4	0.162	0.332	0.039	1.292
5	-0.170	0.114	-0.041	1.251
6	-0.284	0.057	-0.068	1.183
7	-0.341	0.079	-0.082	1.101
8	-0.419	—	-0.101	1.000

Table 3.1-2: Factor Loadings for the UFS-QoL Symptom Severity Scale in TAK-385/CCT-001

Items		Factor1	Factor2	Factor3
Q2	Passing blood clots during your period	**0.763**	0.105	0.073
Q1	Heavy bleeding during your period	**0.759**	0.091	0.123
Q5	Feeling tightness or pressure in pelvis	**0.467**	0.175	0.167
Q8	Feeling fatigued	0.399	0.288	0.078
Q6	Frequent urination in daytime	0.114	**0.965**	0.069
Q7	Frequent nighttime urination	0.212	**0.630**	0.013
Q4	Fluctuation in length of monthly cycle	0.039	0.092	**0.995**
Q3	Fluctuation in duration of menstruation	0.178	0.003	**0.416**

Extraction method: maximum likelihood. Rotation method: orthogonal.

3.2. Development of the Bleeding and Pelvic Discomfort Scale Using Confirmatory Factor Analysis Based on Phase 3 Data

The exploratory factor structure arising from the phase 2 data was assessed using data from the phase 3 studies MVT-601-3001 and MVT-601-3002.

Analyses were based on pooled, blinded data from the first one third of patients enrolled in the two phase 3 studies of relugolix in uterine fibroids (MVT-601-3001 and MVT-601-3002), who completed the patient global assessment of symptoms (PGA) and the Uterine Fibroid Symptom and Health-Related Quality of Life (UFS-QoL) at Baseline and at Week 24.

3.2.1. Confirmatory Factor Analysis using Phase 3 Data

A confirmatory factor analysis was completed using blinded data from one third of phase 3 patients. The acceptance criteria of the confirmatory factor analysis were prespecified as a Goodness of Fit Index > 0.90 and a Comparative Fit Index > 0.90, a root mean square error of approximation < 0.06 and a non-significant p-value to show that the null-hypothesis that the data fits the three-factor model was not rejected (Table 3.2-1).

Factor loadings for the seven-item three-factor model supported the three-factor solution proposed by the exploratory factor analysis in the above described analyses using phase 2 data. Results indicated that the three-factor model, excluding item 8, had a Goodness of Fit Index and a Comparative Fit Index of 1.00 and a root mean square error of approximation of 0.00 (90% CI = 0.00-0.02). The test of model fit returned a p-value of 0.9394. The null hypothesis that the data fit the model was not rejected (see Table 3.2-1). Under this model, Item 5 (Feeling Tightness or Pressure in Pelvis) also cross-loaded onto Factor 2, assessing urinary symptoms.

Table 3.2-1: Confirmatory Factor Analysis of the UFS-QoL Symptom Severity Scale without Item 8: Model Fit Statistics at Baseline (MVT-601-3001 and -3002)

Model Fit Statistics[a]					
Model		**CFI**	**RMSEA (90%CI)**	**GFI**	**P-value**
3-Factor Model (7-item)		1.000	0.000 (0.00-0.02)	1.000	0.9394
Factor Loading[b]					
			Factor1	**Factor2**	**Factor3**
Q1	Heavy bleeding during your period		**0.7314**	0.2672	0.2024
Q2	Passing blood clots during your period		**0.7620**	0.1503	0.2099
Q3	Fluctuation in duration of menstruation		0.3263	0.1861	**0.6909**
Q4	Fluctuation in length of monthly cycle		0.1689	0.1561	**1.0323**
Q5	Feeling tightness or pressure in pelvis		**0.4644**	**0.4657**	0.1965
Q6	Frequent urination in daytime		0.2503	**0.7727**	0.1300
Q7	Frequent night time urination		0.1553	**0.8605**	0.1538

Abbreviations: CFI, comparative fit index; CI, confidence interval; GFI, goodness of fit index; RMSEA, root mean square error approximation.
[a] Model fit statistics allow for assessment of the model appropriateness.
[b] Rotation Method: Orthogonal.

In order to further assess the performance of the Fatigue item, which was excluded following the exploratory factor analysis due to cross-loading, the confirmatory factor analysis was reconducted with the inclusion of this item in Factor 1. Results showed that the eight-item three-factor model had a Goodness of Fit Index of 0.996, a Comparative Fit Index of 1.00 and a root mean square error of approximation of 0.00 (90% CI = 0.00-0.05). The test of model fit returned a p-value of 0.8056. However, the results for Item 8 showed a cross-loading of this item at 0.417 on Factor 1 and 0.437 on Factor 2 (Table 3.2-2). This continued cross-loading supports the exclusion of this item in the scoring of any factor (Table 3.2-2).

Table 3.2-2: Confirmatory Factor Analysis of the UFS-QoL Symptom Severity Scale with Item 8 included: Model Fit Statistics at Baseline (MVT-601-3001 and 3002)

Model Fit Statistics[a]					
Model		**CFI**	**RMSEA (90%CI)**	**GFI**	**P-value**
3-Factor Model (8-item)		1.000	0.000 (0.00-0.05)	0.996	0.8056
Factor Loading[b]					
			Factor1	Factor2	Factor3
Q1	Heavy bleeding during your period		**0.732**	0.265	0.211
Q2	Passing blood clots during your period		**0.750**	0.150	0.226
Q3	Fluctuation in duration of menstruation		0.296	0.175	**0.767**
Q4	Fluctuation in length of monthly cycle		0.180	0.167	**0.932**
Q5	Feeling tightness or pressure in pelvis		**0.473**	**0.465**	0.206
Q6	Frequent urination in daytime		0.251	**0.757**	0.137
Q7	Frequent night time urination		0.150	**0.876**	0.156
Q8	Feeling fatigued		**0.417**	**0.437**	0.136

Abbreviations: CFI, comparative fit index; CI, confidence interval; GFI, goodness of fit index; Q, question; RMSEA, root mean square error of approximation.
[a] Model fit statistics allow for assessment of the model appropriateness.
[b] Rotation Method: Orthogonal.

3.3. Classical Test Theory Psychometric Analyses of the Bleeding and Pelvic Discomfort Scale Based on Phase 3 Data

Each of the above-described factor analyses showed that a seven-item three-factor solution was appropriate for the UFS-QoL Symptom Severity Scale. Following this confirmation, blinded psychometric appraisal of the measure was implemented to further understand the performance of the items and subscales of the UFS-QoL Symptom Severity Scale. For the item level analysis, all items were assessed. For subscale level analysis, the analysis was focused, primarily, on the evaluation of the Factor 1 – the Bleeding and Pelvic Discomfort (BPD) Scale. The BPD Scale was selected as the primary focus for further psychometric evaluation, as it presents clinical and patient-reported symptoms proximal to the disease and is associated with high symptom burden experienced by most patients.

Analyses were based on pooled, blinded data from the first one third of patients enrolled in the two phase 3 studies of relugolix in UF (MVT-601-3001 and MVT-601-3002) who completed the PGA for symptoms and the UFS-QoL at Baseline and at Week 24. Of note, for the analyses specified in this section, only data at Baseline and Week 12 were used.

3.3.1. Item Level Analysis of the UFS-QoL Symptom Severity Scale

UFS-QoL Symptom Severity Scale item responses were assessed for floor (highest possible severity) and ceiling effects (lowest possible severity). Overall, the measure showed no ceiling effects (response option 1, Table 3.3-1, demonstrating that the items have scope to capture

patient improvement in disease burden. A greater proportion of patients responded at floor level (response option 5; range =11.15 to 36.15%), which is expected at the start of a clinical trial. All response options for all items were used, showing a good coverage of the range of disease burden. When considering BPD Scale items, all items showed a range of responses that covered the response scale, with over 50% of patients reporting being a (very) great deal distressed by heavy bleeding during menstrual period (Item 1), passing blot clots during menstrual period (Item 2), and feeling of tightness or pressure in the pelvic area (Item 5).

Table 3.3-1: Summary of UFS-QoL Symptom Severity Scale Response at Baseline by Items in MVT-601-3001 and 3002

	Q1 (N = 260)		Q2 (N = 260)		Q3 (N = 260)		Q4 (N = 260)		Q5 (N = 260)		Q6 (N = 260)		Q7 (N = 260)		Q8 (N = 260)	
Response	n	(%)	n	(%)	n	(%)	n	(%)	n	(%)	n	(%)	n	(%)	n	(%)
1	4	(1.54%)	4	(1.54%)	44	(16.92%)	63	(24.23%)	21	(8.08%)	48	(18.46%)	54	(20.77%)	13	(5.00%)
2	15	(5.77%)	30	(11.54%)	48	(18.46%)	37	(14.23%)	24	(9.23%)	35	(13.46%)	53	(20.38%)	21	(8.08%)
3	53	(20.38%)	61	(23.46%)	66	(25.38%)	69	(26.54%)	57	(21.92%)	77	(29.62%)	64	(24.62%)	59	(22.69%)
4	101	(38.85%)	71	(27.31%)	64	(24.62%)	62	(23.85%)	96	(36.92%)	62	(23.85%)	55	(21.15%)	82	(31.54%)
5	87	(33.46%)	94	(36.15%)	38	(14.62%)	29	(11.15%)	62	(23.85%)	38	(14.62%)	34	(13.08%)	85	(32.69%)

Abbreviations: N, number of patients; n, number of patients in subset; Q, question.

3.3.2. Scale Level Analysis of the BPD Scale

3.3.2.1. Internal Consistency

Internal consistency was assessed for the BPD Scale at Baseline and Week 12. Reliability was acceptable at Baseline (> 0.70) and good at Week 12 (> 0.80; Table 3.3-2).

Table 3.3-2: Cronbach's Alpha Coefficient of BPD Scale by VISIT (MVT-601-3001 and 3002)

		Q1	Q2	Q3	
	n	Mean (SD)	Mean (SD)	Mean (SD)	**Alpha[a]**
Baseline	260	3.97 (0.95)	3.85 (1.09)	3.59 (1.18)	0.768
Week 12	258	2.75 (1.47)	2.69 (1.46)	2.64 (1.36)	0.882

Abbreviations: n, number of patients; Q, question; SD, standard deviation.
[a] Cronbach Coefficient Alpha

3.3.2.2. Item-to-Total Correlations

Item-to-total correlations were assessed to ensure that each item was associated with the BPD Scale score. Correlations demonstrate that each of the items have a strong relationship with the total score at Baseline and at Week 12 ($r > 0.50$) (Table 3.3-3). Correlations improved at Week 12, which represents a greater spread of the data across each item's five-point response scale, further supporting the relationship of these items to the BPD total score.

Table 3.3-3: Intercorrelation of Items in BPD Scale by Visit (MVT-601-3001 and 3002)

Question	**Baseline N = 260**	**Week 12** N = 258
Q1	0.670	0.802
Q2	0.620	0.845
Q5	0.533	0.674

Note: Intercorrelation calculated using Pearson's correlations.

3.3.2.3. Item Discrimination Indices

An item discrimination index was employed to assess the ability of each item to discriminate between high and low severity patients. At Baseline, the discrimination index represents each item's ability to differentiate patients on the BPD Scale scores at a single time point, and at Week 12, the discrimination index represents the ability to differentiate patients based on their level of change from Baseline to Week 12 in the BPD Scale score.

Results show that all items had a discrimination index above 0.60, demonstrating that BPD Scale items are able to discriminate between high- and low-severity patients both when assessing single time point scores and change over time (Table 3.3-4).

Table 3.3-4: Item Discrimination Index of BPD Scale (MVT-601-3001 and 3002)

	Q1	Q2	Q5
Baseline (n = 260)	0.815	0.954	0.923
Week 12 (n = 258)	0.915	0.986	0.836

Abbreviations: n, number of patients; Q, question.
Note: BPD scale upper/lower ranges**:** Upper = at least 65-point reduction, Lower = at most 10-point reduction.

3.3.2.4. Known-Groups Validity

A known-groups analysis assessed the descriptive BPD score and score ranges for patients stratified by level of severity reported on the PGA (symptoms). Results from the known-groups validity assessment show that mean and median BPD Scale scores increase monotonically in line with PGA symptom severity (Table 3.3-5).

3.3.2.5. Ability to Detect Change

The BPD Scale's ability to detect change was assessed though the difference in BPD Scale scores over time in patients who have changed with respect to the measurement concept as measured by the PGA (symptoms). For each PGA stratified group, within person change from Baseline to Week 12 and standardized effect size statistics (SES) for change over the same period were assessed. SES statistics judged were based on Cohen's recommendations (small change, 0.20; moderate change, 0.50; large change, 0.80).

Results showed that the mean change for improving PGA categories had a monotonically increasing pattern from patients who had a PGA change of 0 to patients who had a PGA improvement of -4 (Table 3.3-6). Worsening groups (PGA change of +1 or +2) had very low levels of mean change, with wide standard deviations around the mean due to the low sample size in these categories.

In line with expectations, the SES statistics for the improvement categories (PGA score change of -1 to -4) were large (> 0.80) compared to the moderate SES found in the patients who reported no change (PGA score change of 0; SES = 0.55).

Table 3.3-5: Summary Statistics of BPD Scale Score at Baseline by PGA (symptoms) Response (MVT-601-3001 and 3002)

	Baseline BPD Scale Score[a]						
Baseline PGA	**N**	**Mean**	**SD**	**Median**	**Q1, Q3**	**Min**	**Max**
1	7	53.57	28.81	58.33	25.00, 75.00	16.67	91.67
2	21	59.92	26.56	58.33	41.67, 75.00	8.33	100.00
3	96	62.33	21.18	66.67	41.67, 75.00	8.33	100.00
4	89	75.09	19.48	75.00	66.67, 91.67	16.67	100.00
5	47	83.51	16.53	91.67	75.00, 100.00	41.67	100.00

Abbreviations: BPD, bleeding and pelvic discomfort; max, maximum; min, minimum; N, number of patients; PGA, Patient Global Assessment; Q1, first quartile; Q3, third quartile; SD, standard deviation.
a Transformed Score.

Table 3.3-6: Summary Statistics of Change from Baseline BPD Scale Score to Week 12 by PGA (symptoms) Change from Baseline (MVT-601-3001 and 3002)

PGA Change Category[a]	N	Mean	SD	95% CI	Median	Q1, Q3	Min	Max	Effect Size[b]
-4	23	-48.19	(42.27)	(-66.47, -29.91)	-66.67	-83.33, 0.00	-100.00	25.00	-2.93
-3	50	-49.33	(33.16)	(-58.76, -39.91)	-54.17	-75.00, -25.00	-100.00	33.33	-2.41
-2	74	-27.70	(30.75)	(-34.83, -20.58)	-25.00	-41.67, 0.00	-91.67	25.00	-1.25
-1	48	-23.09	(28.57)	(-31.39, -14.79)	-16.67	-33.33, -8.33	-100.00	33.33	-1.01
0	39	-10.68	(20.32)	(-17.27, -4.10)	-8.33	-25.00, 0.00	-66.67	33.33	-0.55
1	14	1.79	(19.11)	(-9.25, 12.82)	-4.17	-16.67, 8.33	-16.67	33.33	0.07
2	6	-1.39	(29.54)	(-32.39, 29.61)	-12.50	-25.00, 16.67	-25.00	50.00	-0.05

Abbreviations: BPD, blood and pelvic discomfort; CI, confidence interval; max, maximum; min, minimum; N, number of patients; PGA, Patient Global Assessment; Q1, first quartile; Q3, third quartile; SD, standard deviation.
Note: Statistics calculated using transformed score of BPD scale.
[a] The PGA is a five-point, single item patient-reported outcomes tool that measures patient's symptoms. The PGA change category with -4 = Marked Improvement; 0 = No Change, +4 = Markedly Worse.
[b] Standardized effect sizes are calculated as the mean divided by the standard deviation.

3.4. Conclusions

The exploratory factor analysis offered support for a three-factor solution, which included factors assessing Bleeding and Pelvic Discomfort, Urinary Symptoms, and Fluctuation in Menstruation. The Fluctuations in Menstruation factor had an eigenvalue < 1 but had items that loaded at greater than 0.40 and made theoretical sense as a construct.

The exploratory factor analysis showed that Item 8, measuring fatigue, cross-loaded on two factors (Bleeding and Pelvic Discomfort and Urinary Symptoms). Since fatigue is a multidimensional concept that can assess impacts and/or symptoms concurrently, it was not included in the final factor structure. Confirmatory factor analysis on the seven-item three-factor solution provided support for the exploratory factor structure; however, Item 5 cross-loaded between the BPD and Urinary Symptoms factors in this analysis. As Item 5 (Feeling Tightness or Pressure in Pelvis) is a proximal symptom of uterine fibroids, this item was retained as part of the BPD factor.

To ensure that fatigue was not being inappropriately excluded from the three-factor structure, an additional confirmatory factor analysis was conducted with fatigue included within the BPD factor. The inclusion of fatigue in this model continued to show the expected cross-loading of this item. This analysis confirmed that the multidimensional concept of fatigue was not suitable for inclusion in the BPD factor.

The BPD factor, which assesses symptomology most proximal to the disease, was further assessed through classical test theory psychometric evaluation. The results showed that the items of the BPD Scale work cohesively to inform the total score of the measure, and adequately distinguish between severities. At a score level, descriptive statistics were able to support the construct validity and responsiveness of the BPD Scale through showing a monotonic improvement in BPD Scale score in line with patient self-reported improvement on the PGA (symptoms). Additionally, by showing that the items of the BPD Scale perform well together, the psychometric results help to further support the inclusion of the cross-loading Item 5 on the BPD Scale.

APPENDIX 4. APPROACH TO ESTIMATING THE RESPONDER THRESHOLD OF THE UFS-QOL BLEEDING AND PELVIC DISCOMFORT SCALE

The Uterine Fibroid Symptom and Health-Related Quality of Life Bleeding and Pelvic Discomfort (UFS-QoL BPD) Scale includes the following items:

During the previous 3 months, how distressed were you by:

- Heavy bleeding during your menstrual period;
- Passing blood clots during your menstrual period;
- Feeling tightness or pressure in your pelvic area.

Response options include:

- Not at all;
- A little bit;
- Somewhat;
- A great deal;
- A very great deal.

The summary score of the three items included in the UFS-QoL BPD Scale ranges from 0 to 100, where a higher score indicates a higher level of distress and a lower score indicates a lower level of distress.

Change from Baseline to Week 24 in the BPD Scale score is an alpha-protected key secondary endpoint of the pivotal studies (MVT-601-3001 and MVT-601-3002) to evaluate the treatment benefit of relugolix + E2/NETA (Group A) compared with placebo (Group C). Additionally, a responder analysis will be performed between the two groups with respect to proportion of patients who have achieved a meaningful reduction from Baseline to Week 24 in BPD Scale score. This appendix describes the approach used to derive the responder threshold, including both the quantitative and supportive qualitative methods and the respective results.

The meaningful change threshold is the smallest reduction in the BPD Scale score that is considered meaningful by patients (Cohen, 1988; Crosby, 2003; Revicki, 2008; Cappelleri, 2014; Coon, 2018). The magnitude of a meaningful change threshold depends on the magnitude of the correlation between the BPD Scale change score and the Patient Global Assessment (PGA) of symptom severity (anchor) change and the variability of change on the BPD Scale by improvement categories on the PGA of symptom severity (described in Section 4.2.2). Several anchor-based methods will be used; however, the primary analysis will be a measure of central tendency for each improvement category (see Section 4.2.3). Anchor-based methods will use data collected on:

- The BPD Scale score at Baseline and Week 24; and
- The PGA of symptom severity score at Baseline and Week 24.

Results from the anchor-based analyses will be supported by qualitative data collected in a patient interview study (MVT-601-037), a sub-study of the phase 3 trials, in which patients from

selected sites in the United States (US) provided feedback on what they considered to be a meaningful change on the BPD Scale and the PGA of symptom severity (described in Section 4.2.4).

4.2. Statistical Analyses Plan for Estimation of the Responder Threshold

4.2.1. Anchor and Its Correlation with UFS-QoL Endpoint

The PGA of symptom severity uses a five-point verbal rating scale and asks the patient:

> "How severe were your uterine fibroids symptoms, such as heavy bleeding over the last four weeks?"
>
> Response options include:
>
> - Not severe;
> - Mildly severe;
> - Moderately severe;
> - Very severe;
> - Extremely severe.

The categorical change from Baseline to Week 24 in PGA of symptom severity score will be derived, leading to nine possible outcomes ranging from +4 (denoting worsening) to -4 (denoting improvement). The change in PGA of symptom severity at Week 24 will be used as the anchor (see Table 4.2-1).

4.2.2. Target Anchor Category

The target anchor category is the anchor category that represents the minimum meaningful change and is used as the starting point to identify potential candidates for a meaningful change threshold. For the two pivotal studies, the target anchor category will be a one-point category improvement on the PGA of symptom severity score (see Table 4.2-1), as this is typically considered as a minimal clinical important difference on a five-point Likert scale.

Table 4.2-1: Change in PGA of Symptom Severity as Anchor

Anchor	Anchor Change Category	Potential Target Anchor Change Category (To Be Used for Estimation of Meaningful Change Threshold)
Change in PGA of symptom severity	-4, -3, -2, -1 (improvement), 0 (same), +1, +2, +3, +4 (worsening)	-1-category change (improvement)

Abbreviations: PGA = patient global assessment.

4.2.3. Anchor-Based Methods

To determine the meaningful change threshold for the reduction in USF-QoL BPD Scale score, the anchor-based analyses described below will be performed.

The category (or point) change in PGA of symptom severity score will be used as the anchor to classify patients into response groups depending on their level of symptom severity change from Baseline to Week 24 (see Table 4.2-1). Uncollapsed, categorical change on the PGA will range from +4 to -4. Collapsed, categorical change will be considered based on the distribution of change categories on the PGA of symptom severity. Usually the collapsing occurs on the tails with extreme worsening (+4) or improvement (-4).

Among the anchor-based analyses described below, the within-group analysis will be primary and other analyses (including between-group analysis) are supportive.

4.2.3.1. Correlation with Anchor

Correlation between the categorical change on the PGA of symptom severity score and the change in the BPD Scale score will be evaluated at Week 24, using blinded pooled data from the first third of the enrolled patients from the two pivotal studies who have completed Week 24 visits and have the corresponding PGA of symptom severity data available (denoted as the "threshold determination analysis set"). Polyserial correlation coefficient will be used with a criteria value of > 0.30 indicating meaningful correlation (Crosby, 2003; Revicki, 2008; Cappelleri, 2014; Coon, 2018).

4.2.3.2. Within-Group Meaningful Change

Magnitude of change from Baseline to Week 24 in BPD Scale score will be calculated within each anchor category group. Changes in BPD Scale scores are negative for symptom reductions and positive for symptom increases.

Descriptive statistics (*n*, mean change, median change, 25th and 75th percentiles, standard deviation [SD], confidence interval [CI], and standardized effect size [SES]) will be reported for the changes in BPD Scale scores by anchor category. The SES will be calculated for each level of anchor category group by dividing the mean change score of BPD Scale from Baseline by the Baseline SD of the anchor category group. The impact of treatment will be judged based on Cohen's recommendations (1988): small change (SES = 0.20), moderate change (SES = 0.50), and large change (SES = 0.80). Significance associated with within-patient change will be evaluated using paired t-tests on the change in BPD Scale score separately for each level of improvement on the anchor.

4.2.3.3. Supportive Analysis of Between Group Meaningful Change Using Analysis of Variance

Analysis of variance (ANOVA) will be used to determine whether a difference in mean change scores from Baseline to Week 24 on the UFS-QoL BPD Scale exists between the categorical change groups (or the collapsed groups, as appropriate). Providing there is a significant change in UFS-QoL BPD Scale scores between the (collapsed) anchor groups, the between-group differences will be explored. Any anchor group with at least 15 patients will be included in this analysis. An anchor group with < 15 patients (usually occurring on the tails with extreme

worsening [+4] or improvement [-4]) will be collapsed with its adjacent group as appropriate. Comparison of the anchor groups of interest between the target anchor (-1 change category) and the "0 change" category will be performed using a t-test. The statistically significant difference on the BPD Scale change scores corresponding to a 1-category change on the PGA of symptom severity can be used as supportive information for estimating the meaningful change threshold.

4.2.3.4. Visualizing Cumulative Distribution Function and Probability Distribution Function Plots by Anchor Category Group

Anchor-based meaningful change will also be evaluated using cumulative distribution function (CDF) plots utilizing the Kernel smoothing for all anchor category groups, based on cumulative change in UFS-QoL BPD Scale scores for all available changes from Baseline to Week 24. Specifically, the CDF plot for each anchor category displays the probability (presented on the y-axis) of patients who have achieved a given absolute change of X or less in BPD Scale score from Baseline to Week 24 for each point change along the range of possible absolute changes (from -100 [maximum reduction] to 0 [no change] to 100 [maximum increase]) expressed on the x-axis.

Similarly, the smooth probability density function (PDF) will also be plotted for each anchor category group over the range of absolute changes in BPD Scale scores. These probabilities are plotted on the y-axis, with the BPD Scale change score on the x-axis.

The CDF and PDF curves are delineated by anchor improvement category (from -4 to +4) displaying the center and separation between the curve for the target anchor group and the curve for the group reporting no change on PGA of symptom severity. It is expected that the CDF curves will not cross between the change category groups (eg, monotonic increase from no change to slightly improved and moderately improved).

4.2.4. Determining a Meaningful Change Threshold Using the Totality-of-Evidence Approach

The meaningful change threshold will be determined using the totality of evidence from the results of above quantitative anchor-based analyses; results from the interview study (MVT-601-037) will be used as supportive evidence.

The results of these analyses and proposed thresholds will be included into the Patient-Reported Outcome dossier to be submitted at the time of filing.

4.3. Results from Anchor-Based Analyses

4.3.1. Correlation of Change in BPD with PGA of Symptom Severity

Meaningful change for the UFS-QoL BPD Scale was derived based on anchor-based methods, supported by cumulative distribution function (CDF) and probability density function (PDF) curves. To assess the suitability of the selected anchor, PGA of symptom severity, a polyserial correlation was calculated between change on the PGA from Baseline to Week 24 and the change from Baseline to Week 24 on the BPD Scale. The change in the PGA was moderately correlated (r = 0.57) with the change on the BPD Scale (Table 4.3-1). Given that the PGA is less complex than the BPD scale, this result indicates that the PGA is a suitable anchor for the BPD Scale.

4.3.2. Improvement on BPD Scale by PGA Change Category

Uncollapsed changes on the PGA were used to determine minimal meaningful improvement on the BPD Scale (Table 4.3-1). Improvement on the BPD Scale increased monotonically for all the categories from "no change (0)" to "1-category improvement (-1)" to "2-category improvement (-2)" to "3 category improvement (-3)" with nonoverlapping 95% CIs for mean change of the groups. Table 4.3-1 shows further that a 1-category improvement (-1) is associated with a 27.31-point mean improvement in the BPD Scale score at Week 24 compared with Baseline, with a 95% CI [-35.42, -19.19], a large SES = -1.21, and a median improvement of 25.00 points.

Table 4.3-1: Summary of Change from Baseline to Week 24 in UFS-QoL BPD Scale by PGA for Symptom Severity Change Category (mITT Population)

PGA Change Category	N = 255	Change in BPD					Correlation between PGA Change and BPD Change[a]
		Mean (SD)	Median	95% CI	p-value[b]	SES[c]	
4-Category deterioration (+4)	0						0.57
3-Category deterioration (+3)	2	-12.50 (5.89)	-12.5	-65.44, 40.44	0.2048	-2.12	
2-Category deterioration (+2)	2	0.00 (11.79)	0	-105.89, 105.88	1.00	0.0	
1-Category deterioration (+1)	21	-10.32 (16.22)	-8.33	-17.70, -2.93	0.0086	-0.54	
0-Category deterioration (0)	47	-9.93 (23.09)	-8.33	**-16.71**, -3.15	0.005	-0.42	
1-Category improvement (-1)	47	**-27.31** (27.62)	-25.00	-35.42, **-19.19**	< 0.0001	-1.21	
2-Category improvement (-2)	68	-42.16 (25.71)	-41.67	-48.38, -35.93	< 0.0001	-1.93	
3-Category improvement (-3)	45	-61.85 (26.62)	-66.67	-69.85, -53.85	< 0.0001	-3.25	
4-Category improvement (-4)	23	-54.35 (32.65)	-66.67	-68.47, -40.23	< 0.0001	-4.12	

Abbreviations: BPD = bleeding and pelvic discomfort; CI = confidence interval; mITT = modified intent-to-treat; PGA = patient global assessment; SD = standard deviation; SES = standardized effect size.
mITT is used to calculate change from Baseline score at Week 24 and includes patients from the mITT population who have available change from Baseline data at Week 24.
[a] Polyserial correlation coefficient between change in BPD Scale and change in PGA of symptom severity.
[b] The p-value for each individual change group is derived from a paired (within-sample) t-test assessing the difference over time.
[c] SES is calculated as the mean divided by the SD of Baseline. SES is judged as small = 0.2, moderate = 0.5, and large = 0.8 (Cohen 1988).

Table 4.3-2 highlights that the difference between the "1-category improvement" and the "no change" groups (mean = -17.38 with a 95% CI of [-27.81, -6.94]) was statistically significant (p = 0.0013) and had a moderate SES = -0.736, which also supports the notion that patients interpreted these change categories as distinct.

Patients were able to distinguish between the PGA improvement categories, as demonstrated by the nonoverlapping CIs (in Table 4.3-2) for their UFS-QoL BPD Scale scores and as illustrated

by the clear separation between the CDF curves presented in Figure 4.3-1. Since statistically significant differences existed in patient responses on the BPD Scale between the "1-category improvement (-1)" option and the "no change" and "2-category improvement (-2)" options, a 1-category improvement on the PGA was considered a meaningful target anchor category for assessing the responder threshold on the BPD Scale. Although a 2-category improvement could have been considered for deriving the meaningful change threshold, such a threshold would not qualify as being the *minimum* threshold possible. Given the statistical difference between the 1- and 2-category improvements and the fact that patients were able to distinguish between the two response options (to be taken up shortly), the evidence supports using a 1-category improvement on the PGA for estimating the minimum meaningful change threshold. This decision is also supported by qualitative evidence generated from the Exit Interview study (see Section 4.2.4).

Table 4.3-2: Summary of Change from Baseline to Week 24 in BPD Scale Between Target Anchor (-1) and No Change (0) in PGA of Symptom Severity (mITT Population)

Anchor	Categorical Change	N	Mean Change from BL	SD	95% CI	p-value[a]	Baseline SD	SES
PGA	1-category improvement (-1)	47	-27.31	27.62	-35.42, -19.19		22.63	
	No change (0)	47	-9.93	23.09	-16.71, -3.15		23.61	
	Difference		-17.38	25.46	-27.81, -6.94	0.0013		-0.736[b] -0.790[c]

Abbreviations: ANOVA = analysis of variance; BL = Baseline; BPD = bleeding and pelvic discomfort; CI = confidence interval; mITT = modified intent-to-treat; PGA = patient global assessment; SD = standard deviation; SES = standardized effect size.

[a] The p-value is based on t-test for difference in mean change in BPD score between the 2 anchor groups (-1 and 0) from the ANOVA in which the +2, +3, and +4 groups were collapsed with the +1 group due to 0 or few patients in the respective groups.

[b] SES is calculated as the mean difference divided by the SD of Baseline for no change group. They are judged as small = 0.2, moderate = 0.5, and large = 0.8 (Cohen 1988).

[c] SES calculated as the mean difference divided by the standard deviation of Baseline for pooled from all categories (Glass 1976).

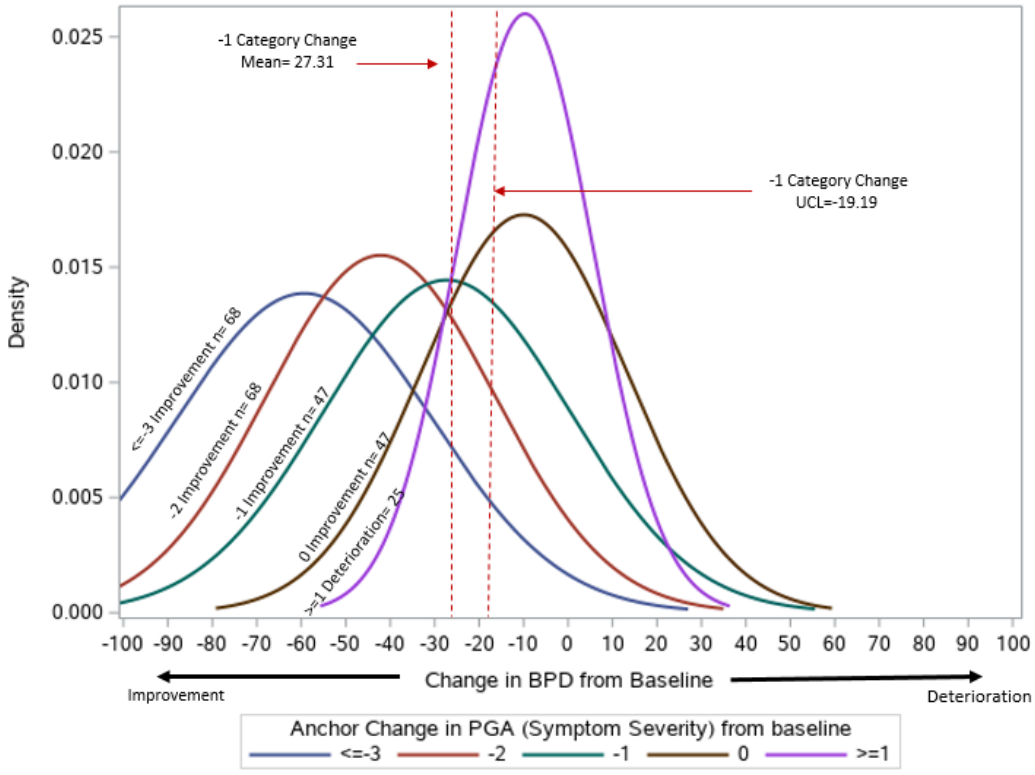
4.3.3. Estimation of Responder Threshold

Examination of the PDF curves, presented in Figure 4.3-1, indicates that the dispersion is roughly the same for the options between "> 3 category improvement" and "no change." The crossing of the "no change" and "1-category improvement" PDF curves at approximately -24 points (ie, a 24-point improvement on the BPD between Baseline and Week 24) indicates the meaningful change threshold is greater (less negative) than this value, because to the left of the value the "1-category improvement" was more probable than the "no change" curve. That is, to the left of this point (larger improvements) patients were more likely to be responders than to the right of this point. However, since the goal is to establish the minimum meaningful change threshold, the value -24 points is likely too conservative.

Using the mean or median values for measuring improvement in the BPD Scale would also yield estimates that are too conservative, because expected values do not necessarily constitute a *minimum* meaningful change threshold for patients. That is, nearly half the patients stratified in

the PGA "1-category improvement" who reported changes smaller than (to the right of) the mean or median on the BPD Scale would be classified as nonresponders by using the mean or median as the threshold despite of their reporting "1-category improvement." A less conservative, though still plausible estimate for the minimal meaningful change threshold is the upper bound of the 95% CI for mean change in the "1-category improvement" group. Its use will result in a smaller proportion of patients being classified as nonresponders in change on the BPD Scale than the expected value (ie, the mean). According to the uncollapsed anchor-based analysis (Table 4.3-1), this value is approximately -19 (ie, a 19-point improvement on the BPD Scale between Baseline and Week 24). Selection of this value is supported by the fact that the mean changes are statistically significantly different (Table 4.3-2) between "no change" and "1-category improvement" groups with clear separation of the respective 95% CIs for mean change. Of note, a value as low as -17 could also be selected, since it is less than the lower-bound 95% CI estimate of -16.71 for the "no change" group.

Figure 4.3-1: PDF of the Change in UFS-QoL BPD Scale by PGA Anchor Change Category (Collapsed)

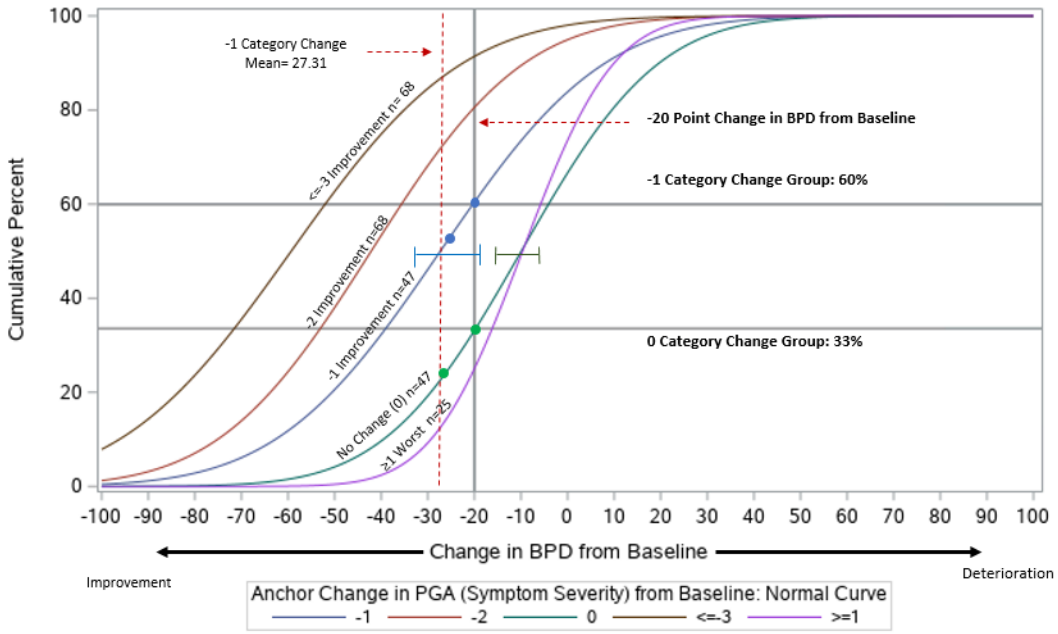


Abbreviations: BPD = bleeding and pelvic discomfort; PGA = patient global assessment; UCL = upper confidence limit.

Examination of the CDF curves for the potential minimum meaningful threshold value of -19 points on the BPD Scale allows one to estimate the cumulative percent of patients that would

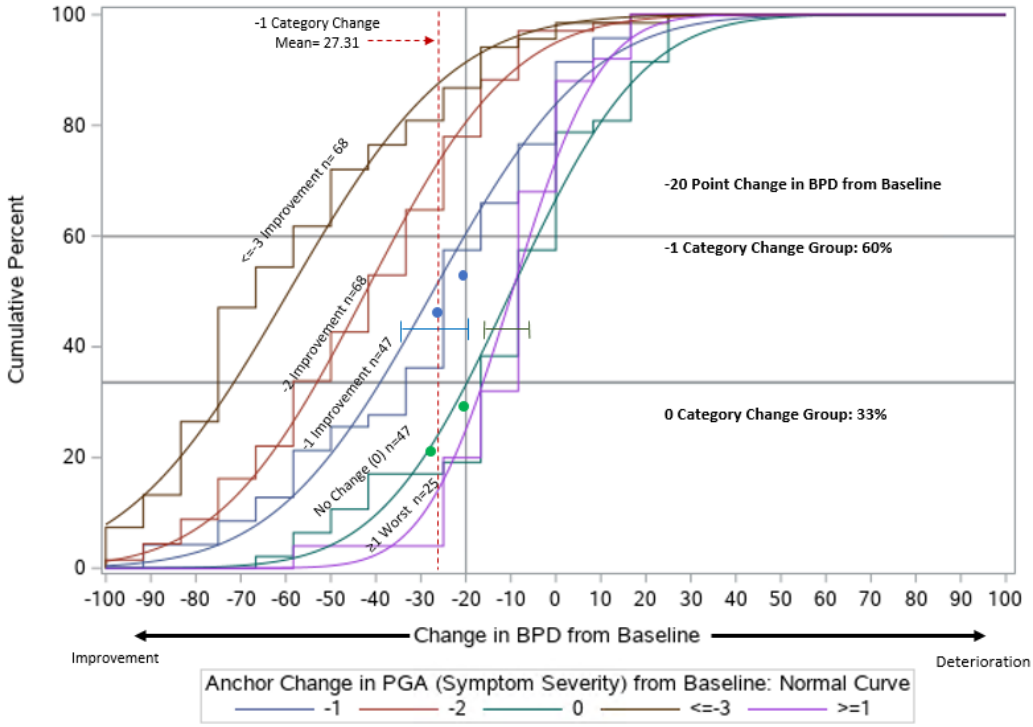
experience the improvement. As illustrated in Figure 4.3-2, approximately 35% of the "no change" group and 61% of the "1-category improvement" group experienced at least a 19-point improvement on the BPD Scale by Week 24. The high percent of patients in the "no change" group who improved on the BPD Scale by Week 24 indicates that setting the minimum meaningful change threshold at 19 points may be too liberal. The percent of misclassified responders can be improved by selecting a slightly larger value. Setting the minimum meaningful change threshold at 20-point improvement on the BPD Scale would decrease slightly the percent of misclassified responders for the "no change" group to 33% while decreasing slightly the percent of patients classified as responders to 60% for the "1-category improvement" group. As supportive information, the empirical CDFs were step-curves (reflecting the discrete nature of the BPD scores) are provided (Figure 4.3-3), indicating that smooth curves are reasonably close to the empirical CDFs.

Figure 4.3-2: Cumulative Distribution Function of Change at Week 24 in UFS-QoL BPD Scale Score by PGA Anchor Change Category (Collapsed)



Abbreviations: BPD = bleeding and pelvic discomfort; PGA = patient global assessment.

Figure 4.3-3: Empirical Cumulative Distribution Function of Change at Week 24 in UFS-QoL BPD Scale Score by PGA Anchor Change Category (Collapsed)



Abbreviations: BPD = bleeding and pelvic discomfort; PGA = patient global assessment.

4.4 Exit Interview Study Synthesis

4.4.1 Objectives

The objectives of the exit interviews were to: 1) provide qualitative evidence to understand meaningful change for patients following clinical intervention and 2) to elicit data on what patients consider to be a minimum meaningful improvement on different patient-reported outcomes (PROs), including:

- The UFS-QoL BPD Scale,
- The PGA symptoms severity.

These objectives were achieved through conducting web/Internet-based video or telephone interviews with English-speaking patients in the US within 3 to 14 days after their Week 24 visit of either ongoing phase 3 clinical study (MVT-601-3001 [LIBERTY 1] or MVT-601-3002 [LIBERTY 2]).

Minimum meaningful improvements on other PROs were also explored as part of the exit interview study; results of the respective exercises will be included in the full report for this exit interview study.

4.4.2 Methodology – Qualitative Interviews

The exit interviews were conducted via a web/Internet-based video platform (Doxy.me [https://doxy.me/]) or via telephone by trained and experienced Endpoint Outcomes interviewers.

In the event that a patient did not improve by at least 1 point from Baseline Day 1 to Week 24 based on her PGA of symptom severity scores, meaningful change exercises were not conducted for any of the PROs. An improvement on the PGA of symptom severity was required so that patients could provide contextually relevant feedback related to positive changes in uterine fibroid symptoms, as they would have experienced an improvement throughout the trial. Table 4.4-1 summarizes the measures/scales of interest, the type of data that was used in the respective meaningful change exercises, and the criteria that must have been met in order for the patient to participate in the respective meaningful change exercise.

Table 4.4-1: Overview of Procedures for Meaningful Change Exercises

Measure/Scale	Type of Data Used	Criteria That Must Have Been Met in Order to Conduct the Respective Meaningful Change Exercise
UFS-QoL BPD Scale (calculated)	MVT-601-3001 (LIBERTY 1) or MVT-601-3002 (LIBERTY 2) Baseline Day 1 response	Improvement on PGA of symptom severity from Baseline Day 1 to Week 24
PGA of symptom severity	MVT-601-3001 (LIBERTY 1) or MVT-601-3002 (LIBERTY 2) responses (Baseline Day 1 and Week 24)	Improvement on PGA of symptom severity from Baseline Day 1 to Week 24

Abbreviations: PGA = patient global assessment; UFS-QoL BPD = Uterine Fibroid Symptom and Health-Related Quality of Life bleeding and pelvic discomfort.

For the UFS-QoL BPD Scale, only patients' clinical study (ie, MVT-601-3001 [LIBERTY 1] or MVT-601-3002 [LIBERTY 2]) Baseline Day 1 data were used during interviews; the meaningful change discussions were hypothetical as Week 24 data were not made available to Endpoint Outcomes.[1] For the UFS-QoL BPD Scale, patients were provided with both their Baseline item-level scores and the summary score calculated based on the three items in the scale. Patients were also given a copy of the three items that comprise the UFS-QoL BPD Scale for reference during the meaningful change exercise. Patients were then presented with prespecified point change increments (ie, 10 points) and asked whether those changes reflected a meaningful improvement. If a patient indicated that a 10-point increment change would be meaningful, she was asked if an increment 5 points fewer would still be meaningful. Using a stepwise approach, interviewers then moved along the scale to identify the point at which minimum meaningful improvement was achieved for the respective patient.

For the PGA of symptom severity, patients were presented with their clinical study scores at Baseline Day 1 and Week 24 and asked if the change was meaningful. Next, patients were presented with a series of hypothetical point changes (ie, more change if the change was not

[1] For secondary endpoint data, only Baseline responses were shared with Endpoint Outcomes.

meaningful or less change if the change was meaningful, as warranted) and asked if those would be meaningful. This process continued until the minimum meaningful change on the PGA of symptom severity for that patient was identified.

Audio recordings of the interviews were transcribed verbatim and anonymized by removing identifying information such as names and places. Each transcript was considered a unit of analysis, and data from all transcripts were aggregated following coding. An initial coding scheme was developed based on the semistructured interview guide and research objectives. The coding scheme was applied and operationalized using Atlas.ti version 8.2.30 (Atlas.ti GmbH, Berlin), a software program designed specifically for qualitative data analysis. Specifically, codes were applied to selected text within each transcript and then queried for frequency across transcripts. Frequencies of patients' interview responses (eg, minimum meaningful change responses) are reported. Minimum meaningful point change medians and ranges were calculated in Excel. As the sample size for the study was small and to reduce the influence of potential outliers, the median is the preferred measure of central tendency reported.

4.4.3 Results

Thirty patients with heavy menstrual bleeding associated with uterine fibroids participated in exit interviews. The average age of these patients was 44, with ages ranging from PPD More than half of the patients (n = PPD self-reported as PPD and most patients (n = PPD were PPD In addition, the majority of patients (n = 26, 86.7%) self-reported some college or higher education as their highest education level. Two patients selected "Other" as the highest level of education and self-reported that they had medical assistant credentials.

The demographic characteristics of the patients from this exit interview study closely matched those of the LIBERTY 1 (MVT-601-3001) and LIBERTY 2 (MVT-601-3002) total sample and the LIBERTY 1 and 2 US sample (see Table 4.4-2). The average age for both the LIBERTY 1 and 2 total sample and US sample was approximately 42 years. Approximately half of participants (n = 396, 51.4%) in the total sample self-reported as black or African American, and over half of the US sample (n = 372, 63.9%) self-reported as black or African American. Additionally, most participants in both the total sample (n = 588, 76.4%) and US sample (n = 450, 77.3%) self-reported as not Hispanic or Latino. Highest level of education data was collected during patient interviews by Endpoint Outcomes; therefore, education level data for all LIBERTY 1 and 2 patients are not available.

Table 4.4-2 includes demographic data for the interviewed study sample as well as the totality of LIBERTY 1 and 2 and the US-based LIBERTY 1 and 2 sample (based on a database snapshot as of 26 Apr 2019).

Table 4.4-2: Patient Demographic Information (from Baseline MVT-601-3001 [LIBERTY 1] or MVT-601-3002 [LIBERTY 2]) and Education Information Collected during Patient Interviews

Baseline Characteristics	Exit Interview Study Sample (N = 30)	LIBERTY 1 and 2 Total Sample (N = 770)	LIBERTY 1 and 2 US Sample (N = 582)
Age (years)			
Mean (SD)	43.9 (4.5)	42.0 (5.4)	42.1 (5.2)
Range	PPD		
Race			
Black or African American	PPD	396 (51.4%)	372 (63.9%)
White		329 (44.4%)	183 (31.4%)
Ethnicity			
Not Hispanic/Latino	PPD	588 (76.4%)	450 (77.3%)
Hispanic/Latino		174 (22.6%)	130 (22.3%)
Highest level of education			
High school (no degree) or less	2 (6.7%)		
High school graduate	2 (6.7%)		
Some college (no degree)	11 (36.7%)		
Associate's degree	4 (13.3%)		
Bachelor's degree	5 (16.7%)		
Master's degree	4 (13.3%)		
Other	2 (6.7%)		

Abbreviations: SD = standard deviation.

Table 4.4-3 below summarizes the total number of exit interview study patients who completed each meaningful change exercise based on the required criteria.

Table 4.4-3: Summary of the Total Number of Exit Interview Study Patients Who Completed Each Meaningful Change Activity

Measure/Scale	Number of Exit Interview Study Patients Participating in Each Exercise (Total N = 30)[2]	Criteria that Must Have Been Met in Order to Conduct the Respective Meaningful Change Exercise
UFS-QoL BPD Scale (calculated)	25	Improvement on PGA of symptom severity from Baseline Day 1 to Week 24
PGA of symptom severity	25	Improvement on PGA of symptom severity from Baseline Day 1 to Week 24

Abbreviations: PGA = patient global assessment; UFS-QoL BPD = Uterine Fibroid Symptom and Health-Related Quality of Life bleeding and pelvic discomfort.

UFS-QoL Bleeding and Pelvic Discomfort Scale

Twenty-five patients improved from Baseline Day 1 to Week 24 on the PGA of symptom severity and participated in the UFS-QoL BPD Scale meaningful change exercise. Data for 24 patients were included in the analysis as one patient provided meaningful change exercise information that was not informative and therefore was excluded from the analysis.[3] The median minimum point change considered to be a meaningful improvement was 10 points (n = 24; range = 5 to 80). The majority of patients completing the UFS-QoL BPD meaningful change activity (n = 15, 62.5%) considered a minimum change of 5 points or 10 points as meaningful (Table 4.4-4).

[2] A total of 30 patients completed exit interviews as part of this study; however, not all 30 patients completed each meaningful change exercise as additional criteria were required in order for a patient to complete the meaningful change exercises. The numbers in this table represent the total number of exit interview patients who met the criteria for participation for the specific meaningful change exercises listed.
[3] This patient did not understand how the three items comprising the UFS-QoL BPD led to the generation of her summary score and could not describe the minimum point change needed for meaningful improvement.

Table 4.4-4: UFS-QoL BPD Scale Meaningful Improvement Results

Minimum Point Change Considered to be a Meaningful Improvement	n (%) [N = 24]
5-point change	11 (45.8%)
10-point change	4 (16.7%)
15-point change	2 (8.3%)
20-point change	0 (0.0%)
25-point change	1 (4.2%)
30-point change	1 (4.2%)
35-point change	1 (4.2%)
40-point change	1 (4.2%)
45-point change	2 (8.3%)
80-point change	1 (4.2%)
Overall point change	
Median	10
Range	5 – 80

Patient Global Assessment of Symptom Severity

Twenty-five patients improved by at least 1 point from Baseline Day 1 to Week 24 on the PGA (for symptoms) and participated in the PGA of symptom severity meaningful change exercise. All patients participating in the PGA of symptom severity meaningful change exercise (n = 25, 100.0%) reported that the actual improvement experienced during the clinical study was meaningful to them.

The median minimum point change considered to be a meaningful improvement was 1 point (n = 24; range = 1 to 3); the most frequently reported minimum meaningful improvement reported by patients was a 1-point change (n = 17, 68.0%) (Table 4.4-5).

Table 4.4-5: PGA Symptom Severity Meaningful Improvement Results

Minimum Point Change Considered to Be a Meaningful Improvement	n (%) [N = 25]
1-point change	17 (68.0%)
2-point change	7 (28.0%)
3-point change	1 (4.0%)
Overall point change	
Median	1
Range	1 – 3

4.4.4 Discussion

The exit interviews provided supportive qualitative evidence to assist in the interpretation of meaningful change in patients following clinical intervention. Patients were required to improve by at least 1 point on the PGA of symptom severity over the course of the clinical study to ensure that patients interviewed had experienced improvement and could reflect upon meaningful improvements in uterine fibroid symptoms.

The decision to use actual clinical trial data in the qualitative interviews was guided by an effort to increase the contextual relevance of each of the meaningful change activities. Providing patients with their Baseline scores for the three PROs created a unique opportunity for patients to reflect on their experience since starting treatment, thereby making the exercises more relevant to them. Further, participation in the meaningful change exercises was predicated on experiencing an improvement in uterine fibroid symptoms over the course of the study, which ensured that patients could speak to meaningful changes stemming from their personal experience. This was confirmed, as all patients participating in the PGA of symptom severity meaningful change exercise (n = 25, 100.0%) reported that the change during the trial was meaningful to them.

These qualitative findings provide patient insight which can be used to supplement psychometric analyses to determine target anchor categories (for the PGA of symptom severity) and responder definitions for the UFS-QoL BPD Scale.

4.5. Determination of Responder Threshold via Triangulation of Findings

Based on the analyses of individual patients' changes in BPD Scale scores, anchored by changes in their response to the PGA of symptom severity, a 20-point change is recommended as the minimum meaningful change threshold for defining a responder. This threshold estimation used the "1-category improvement" PGA group as the target anchor, which is a significantly separated from the "no change" group with respect to the mean change on the BPD Scale. The choice of "1-category improvement" as the target anchor is supported by the majority (17/25, 68%) of the interviewed patients in the exit interview study reporting that a 1-category improvement on the PGA of symptom severity is meaningful to them. The responder threshold of a 20-point change

on the BPD Scale score is larger than what the majority of patients in the exit interview study reported to be meaningful to them, ie, an improvement between 5- to 15-points.

In summary, based on the triangulation of findings from the anchor-based analyses supported by patients' feed-back during exit interviews, a 20-point change in the BPD Scale is proposed as the responder threshold for change in BPD Scale.

4.6. References

Cappelleri JC, Bushmakin AG. Interpretation of patient-reported outcomes. *Statistical methods in medical research* 2014;23:460-483.

Cohen J. Statistical power analysis for the behavioral sciences (1988, 2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

Coon CD, Cook KF. Moving from significance to real-world meaning: methods for interpreting change in clinical outcome assessment scores. *Quality of Life Research* 2018;27:33-40.

Crosby RD, Kolotkin RL, Williams GR. Defining clinically meaningful change in health-related quality of life. *Journal of clinical epidemiology* 2003;56:395-407.Wyrwich KW, Norquist JM, Lenderking WR, Acaster S. Industry Advisory Committee of International Society for Quality of Life R. Methods for interpreting change over time in patient-reported outcome measures. *Qual Life Res* 2013;22:475-483.

Revicki D, Hays RD, Cella D, Sloan J. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *Journal of clinical epidemiology* 2008;61:102-109.

APPENDIX 5. ESTIMATION OF RESPONDER THRESHOLD FOR THE UFS-QOL REVISED ACTIVITIES SCALE

5.1. Approach to Estimating the Responder Threshold of the Revised Activities Scale

The Uterine Fibroid Symptom and Health-Related Quality of Life (UFS-QoL) Revised Activities Scale includes five of the seven most relevant items pertaining to physical and social activities (Coyne 2018). These are:

During the previous 3 months, how often have your symptoms related to uterine fibroids:

- Interfered with your physical activities?
- Made you decrease the amount of time you spent on exercise or other physical activities?
- Made you feel that it was difficult to carry out your usual activities?
- Interfered with your social activities?
- Caused you to plan activities more carefully?

Response options include:

- None of the time;
- A little of the time;
- Some of the time;
- Most of the time;
- All of the time.

The summary score of the five items ranges from 0 to 100, where a lower score indicates a higher ability to do activities (ie, lower score = good) and a higher score indicates a lower ability to do activities.

Change from Baseline to Week 24 in the Revised Activities Scale score is a secondary endpoint of the pivotal studies (MVT-601-3001 and MVT-601-3002) to evaluate the treatment benefit of relugolix + E2/NETA (Group A) compared with placebo (Group C). Additionally, a responder analysis will be performed between the two groups with respect to the proportion of patients who have achieved a meaningful reduction from Baseline to Week 24 in the Revised Activities Scale.

The approach used to derive the responder threshold for improvement in the Revised Activities Scale is similar to that used for the Bleeding and Pelvic Discomfort (BPD) scale (see details in Appendix 4).

This appendix briefly describes the quantitative and supportive qualitative methods and summarizes the respective analysis results.

The meaningful change threshold is the smallest reduction in the Revised Activities Scale score that is considered meaningful by patients (Cohen, 1988; Crosby, 2003; Revicki, 2008; Wyrwich, 2013; Cappelleri, 2014; Coon, 2018). The magnitude of a meaningful change threshold depends

on the magnitude of the correlation between the change in the Revised Activities Scale score and change in anchor (ie, the Patient Global Assessment [PGA] for function anchor) as well as the variability of change on the Revised Activities Scale by improvement categories on the PGA of symptoms (described in Section 5.2.2). Several anchor-based methods will be used; however, the primary analysis will be a measure of central tendency for each improvement category (see Section 5.2.3). Anchor-based methods will use data collected on:

- The UFS-QoL Revised Activities Scale score at Baseline and Week 24; and
- The PGA of function score at Baseline and Week 24.

Results from the anchor-based analyses will be supported by qualitative data collected in a patient interview study (MVT-601-037), a substudy of the phase 3 trials, in which patients from selected sites in the United States (US) provided feedback on what they considered to be a meaningful change on the Revised Activities Scale and the PGA of function (described in Section 5.4).

5.2. Statistical Analysis Plan for Estimation of the Responder Threshold

5.2.1. Anchor and Its Correlation with UFS-QoL Endpoint

The PGA of function uses a five-point verbal rating scale and asks the patient:

> How much were your usual activities limited by uterine fibroid symptoms such as heavy bleeding over the last 4 weeks?

Response options include:

- No limitation at all
- Mild limitation
- Moderate limitation
- Quite a bit of limitation
- Extreme limitation

The categorical change from Baseline to Week 24 in PGA of function score will be derived, leading to nine possible outcomes ranging from +4 (denoting worsening) to -4 (denoting improvement). The change in PGA of function at Week 24 will be used as the anchor (see Table 5.2-1).

5.2.2. Target Anchor Category

The target anchor category is the anchor category that represents the minimum meaningful change and is used as the starting point to identify potential candidates for a meaningful change threshold. For the two pivotal studies, the target anchor category will be a one-point category improvement on the PGA of function (see Table 5.2-1), as this is typically considered as a minimal clinical important difference on a five-point Likert scale.

Table 5.2-1: Change in PGA as Anchor

Anchor	Anchor Change Category	Potential Target Anchor Change Category (To Be Used for Estimation of Meaningful Change Threshold)
Change in PGA of function	-4, -3, -2, -1 (improvement), 0 (same), +1, +2, +3, +4 (worsening)	-1-category change (improvement)

Abbreviations: PGA = patient global assessment.

5.2.3. Anchor-Based Methods

To determine the meaningful change threshold for the reduction in UFS-QoL Revised Activities Scale score, the anchor-based analyses described below will be performed.

The category (or point) change in PGA of function score will be used as the anchor to classify patients into response groups, depending on their level of change in the Revised Activities Scale from Baseline to Week 24 (see Table 5.2-1). Uncollapsed, categorical change on the PGA will range from +4 to -4. Collapsed, categorical change will be considered based on the distribution of change categories on the PGA of function. Usually, the collapsing occurs on the tails with extreme worsening (+4) or improvement (-4).

Among the anchor-based analyses described below, the within-group analysis will be primary and other analyses (including between-group analysis) are supportive.

5.2.3.1. Correlation with Anchor

Correlation between the categorical change on the PGA of function score and the change in the Revised Activities Scale score will be evaluated at Week 24, using blinded pooled data from the first third of the enrolled patients from the two pivotal studies who had completed Week 24 visits and had the corresponding PGA of function data available (denoted as the "threshold determination analysis set"). Polyserial correlation coefficient will be used with a criteria value of > 0.30 indicating meaningful correlation (Cohen, 1988; Crosby, 2003; Revicki, 2008; Cappelleri, 2014; Coon, 2018).

5.2.3.2. Within-Group Meaningful Change

The magnitude of change from Baseline to Week 24 in Revised Activities Scale score will be calculated within each anchor category group. Changes in Revised Activities Scale scores are negative for reduced ability to do activities (indicating a worse outcome) and positive for increased ability to do activities (indicating a better outcome).

Descriptive statistics (*n*, mean change, median change, 25th and 75th percentiles, standard deviation [SD], confidence interval [CI], and standardized effect size [SES]) will be reported for the changes in Revised Activities Scale scores by anchor category. The SES will be calculated for each level of anchor category group by dividing the mean change score of Revised Activities Scale from Baseline by the Baseline SD of the anchor category group. The impact of treatment will be judged based on Cohen's recommendations (1988): small change (SES = 0.20),

moderate change (SES = 0.50), and large change (SES = 0.80). Significance associated within-patient change will be evaluated using paired t-tests on the change in Revised Activities Scale score separately for each level of improvement on the anchor.

5.2.3.3. Supportive Analysis of Between Group Meaningful Change Using Analysis of Variance

Analysis of variance (ANOVA) will be used to determine whether a difference in mean change scores from Baseline to Week 24 on the Revised Activities Scale exists between the categorical change groups (or the collapsed groups, as appropriate). Providing there is a significant change in Revised Activities Scale scores between the (collapsed) anchor groups, the between-group differences will be explored. Any anchor group with at least 15 patients will be included in this analysis. An anchor group with < 15 patients (usually occurring on the tails with extreme worsening [+4] or improvement [-4]) will be collapsed with its adjacent group as appropriate. Comparison of the anchor groups of interest between the target anchor ("-1 change" category) and "0 change" category will be performed using a t-test. A statistically significant difference on the Revised Activities Scale change scores corresponding to a 1-category change on the PGA of function can be used as supportive information for estimating the meaningful change threshold.

5.2.3.4. Visualizing Cumulative Distribution Function and Probability Distribution Function Plots by Anchor Category Group

Anchor-based meaningful change will also be evaluated using cumulative distribution function (CDF) plots utilizing the Kernel smoothing for all anchor category groups, based on cumulative change in the Revised Activities Scale scores for all available changes from Baseline to Week 24. Specifically, the CDF plot for each anchor category displays the probability (presented on y-axis) of patients who have achieved a given absolute change of X or less in the Revised Activities Scale score from Baseline to Week 24 for each point change along the range of possible absolute changes (from -100 [maximum reduction] to 0 [no change] to 100 [maximum increase]) expressed on the x-axis.

Similarly, the smooth probability density function (PDF) will also be plotted for each anchor category group over the range of absolute changes in the Revised Activities Scale scores. These probabilities are plotted on the y-axis with the Revised Activities Scale change score on the x-axis.

The CDF and PDF curves are delineated by anchor improvement category (from -4 to +4) displaying the center and separation between the curve for the target anchor group and the curve for the group reporting no change on PGA of function. It is expected that the CDF curves will not cross between the change category groups (eg, monotonic increase from no change to slightly improved and moderately improved).

5.2.4. Determining a Meaningful Change Threshold Using Totality-of-Evidence Approach

The meaningful change threshold will be determined using the totality of evidence from the results of above quantitative anchor-based analyses; results from the interview study (MVT-601-037) will be used as supportive evidence.

The results of these analyses and proposed thresholds will be included into the Patient-Reported Outcome dossier to be submitted at time of filing.

5.3. Results from Anchor-Based Analyses

5.3.1. Correlation of Change in Revised Activities Scale Score with PGA of Function

Meaningful change for the UFS-QoL Revised Activities Scale was derived based on anchor-based methods, supported by CDF and PDF curves. To assess the suitability of the selected anchor, PGA of function, a polyserial correlation was calculated between change on the PGA from Baseline to Week 24 and the change from Baseline to Week 24 on the Revised Activities Scale. The change in the PGA was moderately negatively correlated (r = -0.60) with the change on the Revised Activities Scale (Table 5.3-1). Given that the PGA of function is less complex than the Revised Activities Scale, this result indicates that the PGA of function is a suitable anchor for the Revised Activities Scale.

Table 5.3-1: Summary of Change from Baseline to Week 24 in UFS-QoL Revised Activities Scale by PGA of Function Change Category (mITT Population)

PGA of Function Change Category	N = 254	Change in Revised Activities					Correlation between PGA Change and Revised Activities Change[a]
		Mean (SD)	Median	95% CI	p-value[b]	SES[c]	
4-category deterioration (+4)	**2**	5.00 (7.07)	5	-58.53,68.53	0.500	0.28	-0.60
3-category deteriorations (+3)	**2**	0	0	-	-	0.00	
2-category deteriorations (+2)	5	7.00 (22.80)	0	-21.31,35.31	0.5302	0.61	
1-category deteriorations (+1)	**22**	-1.59 (23.82)	-5	-12.15,8.97	0.7572	-0.06	
0 Category deteriorations (0)	71	11.55 (28.51)	5	4.80,**18.30**	0.0011	0.38	
1-category improvement (-1)	53	**27.92** (25.65)	**20**	**20.85**,35.00	< 0.0001	1.06	
2-category improvement (-2)	**51**	51.86 (27.60)	60	44.10,59.63	< 0.0001	2.17	
3-category improvement (-3)	**35**	56.81 (27.49)	57.50	47.50,66.11	< 0.0001	2.91	
4-category improvement (-4)	13	60.77 (31.55)	70	41.71, 79.83	< 0.0001	4.40	

Abbreviations: CI = confidence interval; mITT = modified intent-to-treat; PGA = patient global assessment; SD = standard deviation; SES = standardized effect size.
mITT is used to calculate change from Baseline score at Week 24 and includes patients from the mITT population who have available change from Baseline data at Week 24.

[a] Polyserial correlation coefficient between change in Revised Activities Scale and change in PGA of function.

[b] The p-value for each individual change group is derived from a paired (within-sample) t-test assessing the difference over time.

[c] SES calculated as the mean divided by the SD of Baseline. SES is judged as small = 0.2, moderate = 0.5, and large = 0.8 (Cohen 1988).

5.3.2. Improvement on Revised Activities Scale by PGA Change Category

Uncollapsed changes on the PGA of function were used to determine minimal meaningful improvement on the Revised Activities Scale (Table 5.3-1). Improvement on the Revised Activities Scale increased monotonically for all the categories from "no change (0)" to "1-category improvement (-1)" to "2-category improvement (-2)" with non-overlapping 95% CIs for mean change of the three groups. Table 5.3-2 shows that a one category improvement (-1) is associated with a 27.92-point mean improvement in the Revised Activities Scale score at Week 24 compared to Baseline, with a 95% CI [20.85, 35.00], a large SES = 1.06, and a median improvement of 20 points.

Table 5.3-2 highlights that the difference between the "1-category improvement" and the "no change" groups (mean =11.55 with a 95% CI of [4.80, 18.30]) was statistically significant (p = 0.0013) with a moderate SES = 0.54, which reasonably supports the notion that patients interpreted these change categories as distinct.

Table 5.3-2: Summary of Change from Baseline to Week 24 in Revised Activities Scale Between Target Anchor (-1) and No change (0) in PGA of Function (mITT Population)

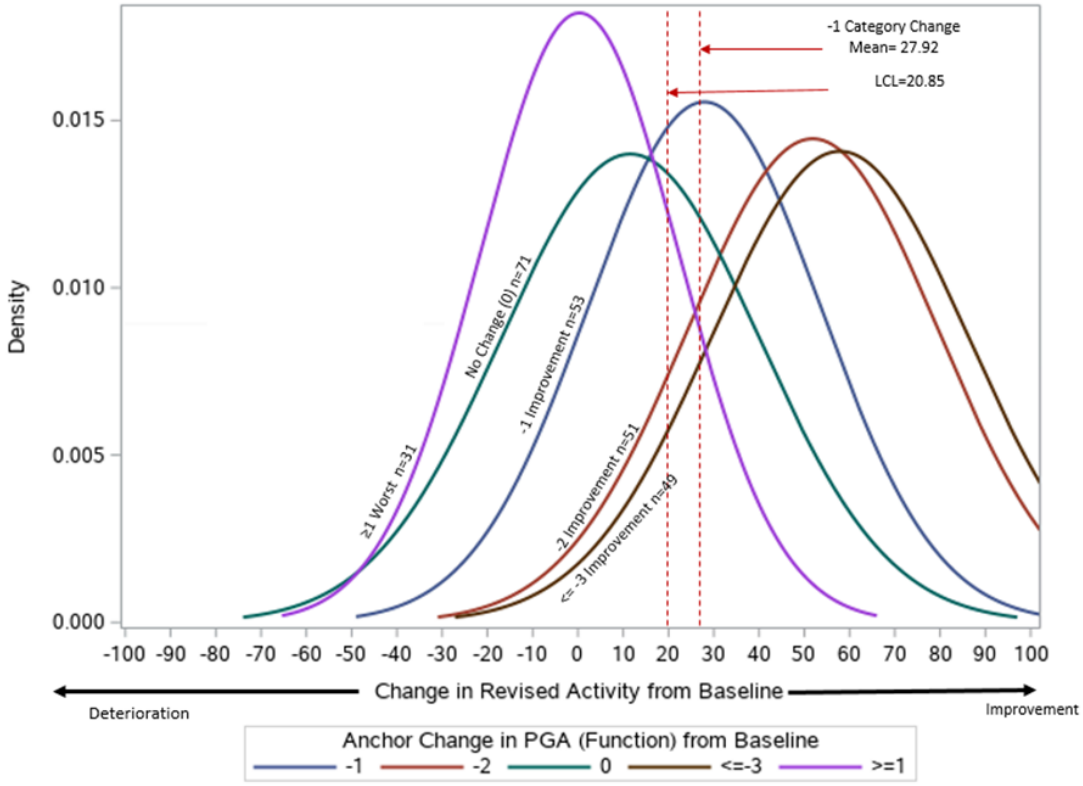
Anchor	Categorical Change	N	Mean Change from BL	SD	95% CI	p-value[a]	Baseline SD	SES
PGA	1-category improvement (-1)	53	**27.92**	**25.65**	**20.85, 35.0**			
	No change (0)	71	**11.55**	**28.51**	**4.80, 18.30**			
	Difference		**16.38**	**27.33**	**6.55, 26.20**	0.0013		0.54[b]
								0.57[c]

[a] The p-value is based on t-test for difference in mean change in BPD score between the 2 anchor groups (-1 and 0) from the ANOVA.
[b] SES calculated as the mean difference divided by the standard deviation of Baseline for no change group. They are judged as small=0.2, moderate=0.5 and large=0.8 (Cohen, 1988).
[c] SES calculated as the mean difference divided by the standard deviation of Baseline for pooled from all categories (Glass, 1976).

That patients were able to distinguish between the PGA "1-category improvement" and the 'no change" group is further supported by the nonoverlapping CIs (in Table 5.3-2) for the respective UFS-QoL Revised Activities Scale scores and as illustrated by the separation between the CDF curves presented in Figure 5.3-1. Since statistically significant differences existed in patient responses on the Revised Activities Scale between the "1-category improvement (-1) " option and the "no change" and the "2-category improvement (-2)" groups, a 1-category improvement on the PGA was considered a meaningful target anchor category for assessing the responder threshold on the Revised Activities Scale. Although a two-category improvement could have been considered for deriving the meaningful change threshold, such a threshold would not qualify as being the *minimum* threshold possible. The evidence (ie, the statistical difference between the1- and 2-category improvements and the fact that patients were able to distinguish between the two response options) supports using a 1-category improvement on the PGA of

function for estimating the minimum meaningful change threshold. This decision is also supported by qualitative evidence generated from the Exit Interview study (see Section 5.4).

Figure 5.3-1: PDF of the Change in UFS-QoL Revised Activities by PGA of Function Anchor Change Category (Collapsed)



Abbreviations: PGA = patient global assessment; LCL = lower confidence limit.

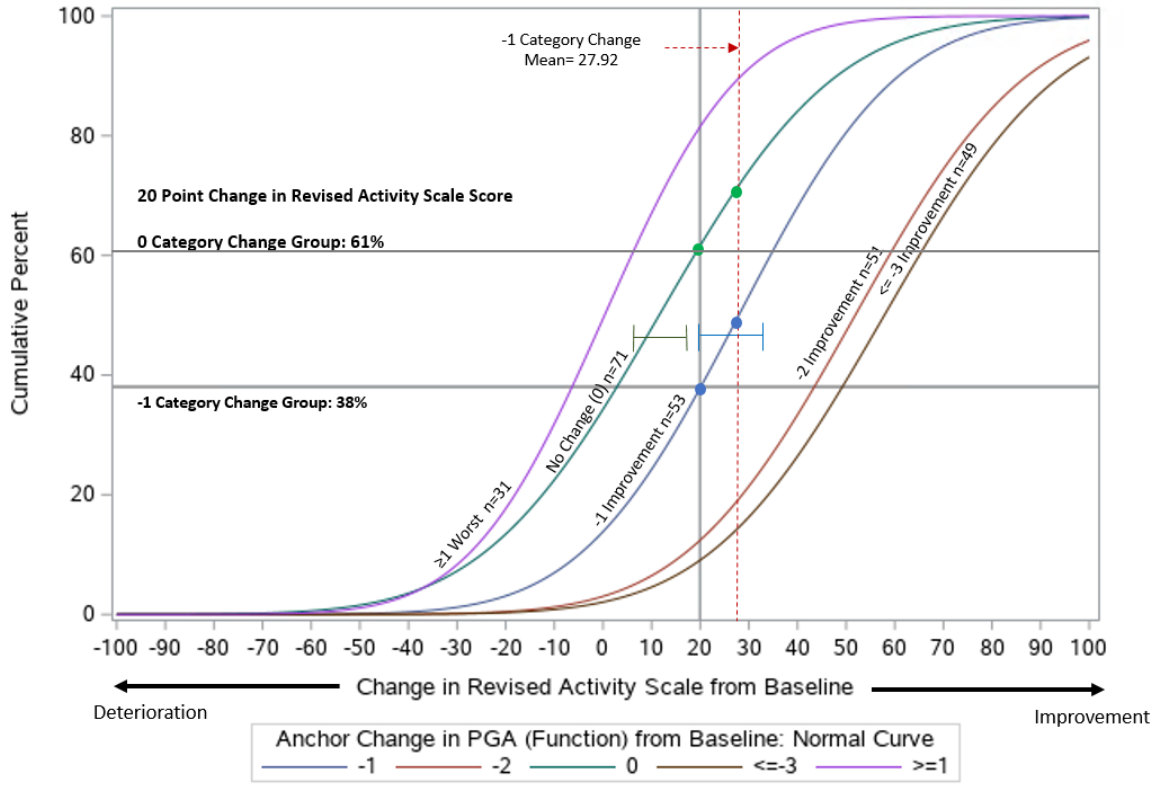
5.3.3. Estimation of Responder Threshold

Using the mean value for measuring improvement in the Revised Activities Scale would yield estimates that are conservative because expected values do not necessarily constitute a *minimum* meaningful change threshold for patients. That is, nearly half the patients stratified in the PGA "1-category improvement" who reported changes smaller than the mean on the Revised Activities Scale would be classified as nonresponders by using the mean as the threshold despite of their reporting "1-category improvement". A less conservative, though still plausible estimate for the minimal meaningful change threshold is the lower bound of the 95% CI for mean change in the "1-category improvement" group. Its use will result in a smaller proportion of patients being classified as nonresponders on the Revised Activities Scale than the expected value (ie, the mean). Similarly, one can also consider the median value since it is less influenced by outliers than either the mean or CI estimates.

According to the uncollapsed anchor-based analysis (Table 5.3-1), the median value for a "1-category improvement" is 20-points, while the lower bound 95% CI for this group is about 21-points (ie, a 21-point improvement on the revised activities between Baseline and Week 24). Given the large discrepancy between the mean and median values suggests that outliers were present in the data; hence, the median value is recommended as a potential minimum change threshold.

Examination of the CDF curves for the potential minimum meaningful threshold value of 20 points on the Revised Activities Scale allows one to estimate the cumulative percent of patients that would experience the improvement. As illustrated in Figure 5.3-2, approximately 38% of the "no change" group and 61% of the "1-category improvement" group experienced at least a 20-point improvement (eg, approximately 62% of the "no change" group and 39% of the "1-category improvement" group experienced less than a 20-point improvement to the left) on the Revised Activities Scale by Week 24.

Figure 5.3-2: Cumulative Distribution Function of Change at Week 24in UFS-QoL Revised Activities Scale Score by PGA Anchor Change Category (Collapsed)

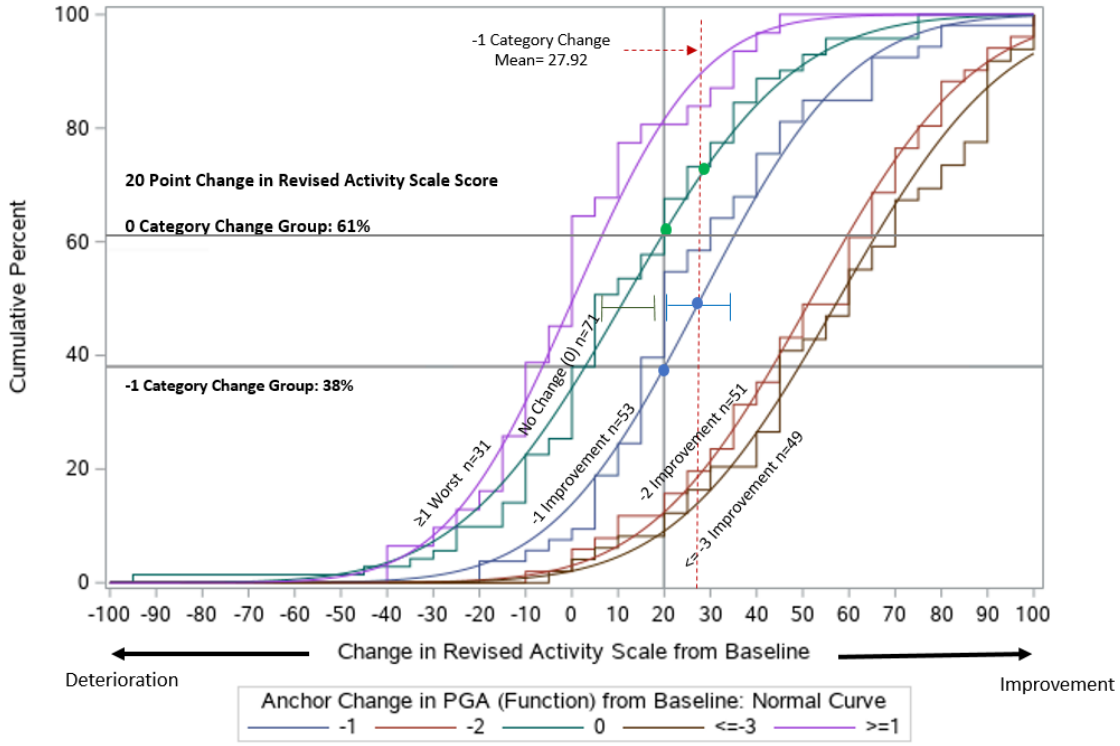


Abbreviations: PGA = patient global assessment.

As supportive information, the empirical CDFs with step-curves (reflecting the discrete nature of the revised activities scores) are provided (Figure 5.3-3), indicating that smooth curves are

reasonably close to the empirical CDFs. Examination of the PDF curves presented in Figure 5.3-1 indicates that the dispersion is roughly the same for the options between "> -3-category improvement" and "no change."

Figure 5.3-3: Empirical Cumulative Distribution Function of Change at Week 24 in UFS-QoL Revised Activities Scale Score by PGA Anchor Change Category (Collapsed)



Abbreviations: PGA = patient global assessment.

5.4. Exit Interview Study Synthesis

5.4.1 Objectives

The objectives of the exit interviews were: 1) to provide qualitative evidence to understand meaningful change for patients following clinical intervention and 2) to elicit data on what patients consider to be a minimum meaningful improvement on different patient-reported outcomes (PROs), including:

- The UFS-QoL Revised Activities Scale;
- The PGA of function.

These objectives were achieved through conducting web/Internet-based video or telephone interviews with English-speaking patients in the US within 3 to 14 days after their Week 24 visit

of either ongoing phase 3 clinical study (MVT-601-3001 [LIBERTY 1] or MVT-601-3002 [LIBERTY 2]).

Minimum meaningful improvements on other PROs were also explored as part of the exit interview study; results of the respective exercises will be included in the full report for this exit interview study.

5.4.2 Methodology – Qualitative Interviews

The exit interviews were conducted via a web/Internet-based video platform (Doxy.me [https://doxy.me/]) or via telephone by trained and experienced Endpoint Outcomes interviewers.

If a patient did not improve by at least 1 point from Baseline Day 1 to Week 24 based on her PGA of function, meaningful change exercises were not conducted for the PGA of function and the UFS-QoL Revised Activities Scale. An improvement on the PGA of function was required so that patients could provide contextually relevant feedback related to positive changes as they would have experienced an improvement throughout the trial. Table 5.4-1 summarizes the measures/scales of interest, the type of data that was used in the respective meaningful change exercises, and the criteria that must have been met in order for the patient to participate in the respective meaningful change exercise.

Table 5.4-1: Overview of Procedures for Meaningful Change Exercises

Measure/Scale	Type of Data Used	Criteria That Must Have Been Met in Order to Conduct the Respective Meaningful Change Exercise
UFS-QoL Revised Activities Scale (calculated)	MVT-601-3001 (LIBERTY 1) or MVT-601-3002 (LIBERTY 2) Baseline Day 1 response	Improvement on PGA of function from Baseline Day 1 to Week 24
PGA (for function)	MVT-601-3001 (LIBERTY 1) or MVT-601-3002 (LIBERTY 2) responses (Baseline Day 1 and Week 24)	Improvement on PGA of function from Baseline Day 1 to Week 24

Abbreviations: PGA = patient global assessment; UFS-QoL = Uterine Fibroid Symptom and Health-Related Quality of Life.

For the UFS-QOL Revised Activities Scale, only patients' clinical study (ie, MVT-601-3001 [LIBERTY 1] or MVT-601-3002 [LIBERTY 2]) Baseline Day 1 data were used during interviews; the meaningful change discussions were hypothetical, as Week 24 data were not made available to Endpoint Outcomes.[4] For the UFS-QoL Revised Activities Scale, patients were provided with both their Baseline item-level scores and the summary score calculated based on the five items in the scale. Patients were also given a copy of the five items that comprise the UFS-QoL Revised Activities Scale for reference during the meaningful change exercise. Patients were then presented with pre-specified point change increments (ie, 10 points) and asked whether those changes reflected a meaningful improvement. If a patient indicated that a 10-point

[4] For secondary endpoint data, only Baseline responses were shared with Endpoint Outcomes.

increment change would be meaningful, she was asked if an increment 5 points fewer would still be meaningful. Using a stepwise approach, interviewers then moved along the scale to identify the point at which minimum meaningful improvement was achieved for the respective patient.

For the PGA of function, patients were presented with their clinical study scores at Baseline Day 1 and Week 24 and were asked if the change was meaningful. Next, patients were presented with a series of hypothetical point changes (ie, more change if the change was not meaningful or less change if the change was meaningful, as warranted) and asked if those would be meaningful. This process continued until the minimum meaningful change on the PGA of function for that patient was identified.

Audio-recordings of the interviews were transcribed verbatim and anonymized by removing identifying information such as names and places. Each transcript was considered a unit of analysis, and data from all transcripts were aggregated following coding. An initial coding scheme was developed based on the semi-structured interview guide and research objectives. The coding scheme was applied and operationalized using Atlas.ti version 8.2.30 (Atlas.ti GmbH, Berlin), a software program designed specifically for qualitative data analysis. Specifically, codes were applied to selected text within each transcript and then queried for frequency across transcripts. Frequencies of patients' interview responses (eg, minimum meaningful change responses) are reported. Minimum meaningful point change medians and ranges were calculated in Excel. As the sample size for the study was small and to reduce the influence of potential outliers, the median is the preferred measure of central tendency reported.

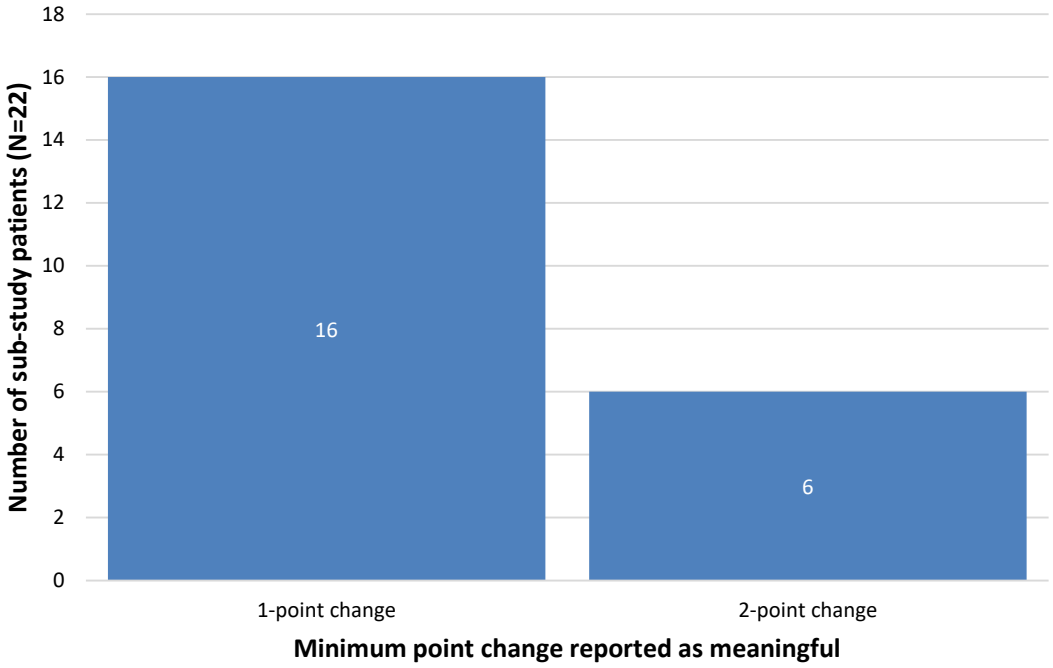
5.4.3 Results

5.4.3.1 PGA of Function[5]

Twenty-two patients improved from Baseline Day 1 to Week 24 on the PGA of function and participated in the PGA of function meaningful change exercise. The demographic characteristics of the 22 patients who completed the PGA of function closely match that of the entire substudy sample as the sample was mostly PPD (n = PPD PPD (n = PPD had completed at least some college or higher (n = 19, 86.4%), and had an average age of approximately 44 years. The median minimum point change considered to be a meaningful improvement was 1 point (n = 22, range = 1-2); the most frequently reported minimum meaningful improvement reported by patients was a 1-point change (n = 16, 72.7%) (Figure 5.4-1).

[5] The PGA of function asks: How much were your usual activities limited by uterine fibroids symptoms such as heavy bleeding over the last 4 weeks? Response options include: No limitation at all, mild limitation, moderate limitation, quite a bit of limitation, and extreme limitation.

Figure 5.4-1: Meaningful Change Estimation: Results of the PGA (for Function)



5.4.3.2 UFS-QoL Revised Activities Subscale[6]

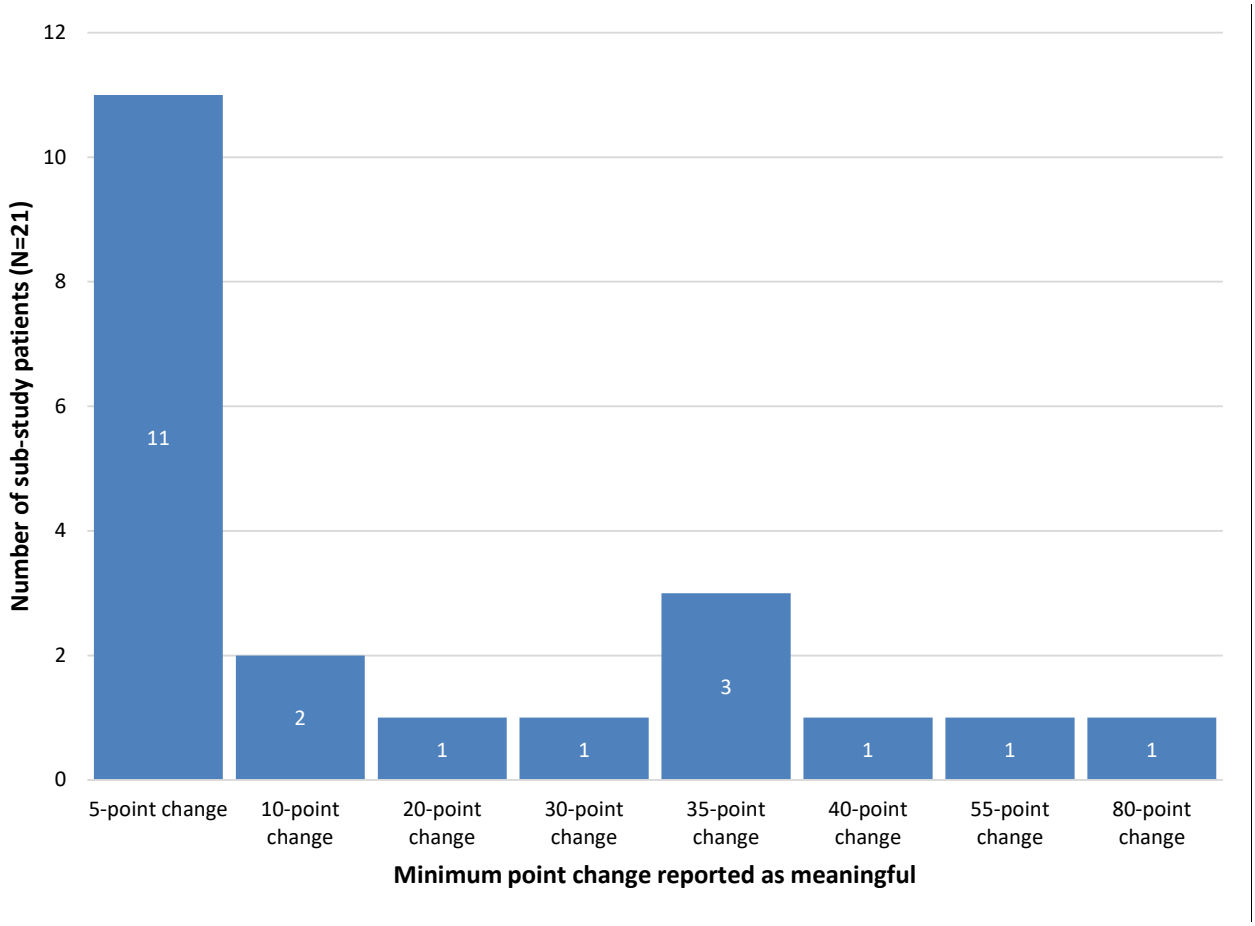
Twenty-two patients improved from Baseline Day 1 to Week 24 on the PGA of function and participated in the UFS-QoL revised activities subscale meaningful change exercise. Data for 21 patients were included in the analysis as one patient provided meaningful change exercise information that was not informative and therefore was excluded from the analysis.[7] The demographic characteristics of the 21 patients who completed the UFS-QoL Revised Activities Scale closely match that of the entire substudy sample as the sample was mostly PPD (n = PPD (n = PPD had completed at least some college or higher (n = 19, 90.5%), and had an average age of approximately 44 years.

[6] The UFS-QoL revised activities subscale includes five items, which ask: During the previous 3 months, how often have your symptoms related to uterine fibroids … 11) interfered with your physical activities; 13) made you decrease the amount of time you spent on exercise or other physical activities; 19) made you feel it was difficult to carry out your usual activities; 20) interfered with your social activities; and 27) made you plan activities more carefully. Response options include 1) None of the time, 2) A little of the time, 3) Some of the time, 4) Most of the time, and 5) All of the time. The score range for the subscale is 0-100. A higher score on the revised activities subscale indicates a lower interference in activities while a lower score on the subscale indicates a higher interference in activities.

[7] This patient was unwilling to describe the minimum point change needed for meaningful improvement for the UFS-QoL revised activity subscale.

The median minimum point change considered to be a meaningful improvement was 5 points (n = 21, range = 5-80); the most frequently reported minimum meaningful improvement reported by patients was a 5-point change (n = 11, 52.4%) (Figure 5.4-2).

Figure 5.4-2: Meaningful Change Estimation: Results of the UFS-QoL Revised Activities Subscale



5.5. Determination of Responder Threshold via Triangulation of Findings

Based on the analyses of individual patient's change in Revised Activities Scale scores anchored by change in their response to the PGA of function, a 20-point change is recommended as the minimum meaningful change threshold for defining a responder. This threshold estimation used the "1-category improvement" PGA group as the target anchor, which is significantly separated from the "no change" group with respect to the mean change on the Revised Activities Scale. The choice of "1-category improvement" as the target anchor is supported by the majority (16/22, 73%) of the interviewed patients in the exit interview study reporting that a 1-category improvement on the PGA of function is meaningful to them. The responder threshold of a 20-point change on the Revised Activities Scale score is larger than what the majority of patients in the exit interview study reported to be meaningful to them (ie, improvements of 5 points [11/21] and 10 points [2/21]).

In summary, based on the triangulation of findings from the anchor-based analyses supported by patients' feedback during exit interviews, a 20-point change in the Revised Activities Scale is proposed as the responder threshold for change in Revised Activities Scale.

5.6. References

Cappelleri JC, Bushmakin AG. Interpretation of patient-reported outcomes. Stat Meth Med Res 2014;23:460-483.

Cohen J. Statistical power analysis for the behavioral sciences (1988, 2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

Coon CD, Cook KF. Moving from significance to real-world meaning: methods for interpreting change in clinical outcome assessment scores. Qual Life Res 2018;27:33-40.

Crosby RD, Kolotkin RL, Williams GR. Defining clinically meaningful change in health-related quality of life. J Clin Epidemiol 2003;56:395-407.

Coyne KS, Harrington A, et al. Psychometric validating of the 1-month recall Uterine Fibroid Symptom and Health-Related Quality of Life Questionnaire (UFS-QOL). ISPOR 23rd Annual International Meeting, 2018

Revicki D, Hays RD, Cella D, Sloan J. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. J Clin Epidemiol 2008;61:102-109.

Wyrwich KW, Norquist JM, Lenderking WR, Acaster S. Industry Advisory Committee of International Society for Quality of Life R. Methods for interpreting change over time in patient-reported outcome measures. Qual Life Res 2013;22:475-483.

STATISTICAL ANALYSIS PLAN

Study Titles:	LIBERTY 1: An International Phase 3 Randomized, Double-Blind, Placebo-Controlled Efficacy and Safety Study to Evaluate Relugolix Co-Administered with and without Low-Dose Estradiol and Norethindrone Acetate in Women with Heavy Menstrual Bleeding Associated with Uterine Fibroids LIBERTY 2: An International Phase 3 Randomized, Double-Blind, Placebo-Controlled Efficacy and Safety Study to Evaluate Relugolix Co-Administered with and without Low-Dose Estradiol and Norethindrone Acetate in Women with Heavy Menstrual Bleeding Associated with Uterine Fibroids
Investigational Product:	Relugolix
Protocol Number:	MVT-601-3001 and MVT-601-3002
Indication:	Heavy menstrual bleeding associated with uterine fibroids
Sponsor:	Myovant Sciences GmbH Viaduktstrasse 8 4051 Basel Switzerland
Regulatory Identifier(s):	IND # 131161 EudraCT # 2016-003727-27
Version/Effective Date:	07-May-2019

STATISTICAL ANALYSIS PLAN APPROVAL SHEET

MVT-601-3001 (LIBERTY 1): An International Phase 3 Randomized, Double-Blind, Placebo-Controlled Efficacy and Safety Study to Evaluate Relugolix Co-Administered with and without Low-Dose Estradiol and Norethindrone Acetate in Women with Heavy Menstrual Bleeding Associated with Uterine Fibroids

MVT-601-3002 (LIBERTY 2): An International Phase 3 Randomized, Double-Blind, Placebo-Controlled Efficacy and Safety Study to Evaluate Relugolix Co-Administered with and without Low-Dose Estradiol and Norethindrone Acetate in Women with Heavy Menstrual Bleeding Associated with Uterine Fibroids

This statistical analysis plan has been approved by Myovant Sciences GmbH ("Myovant"), with Myovant Sciences, Inc., acting as agent of Myovant. The following signatures document this approval.

PPD

07 May 2019
Date

07 May 2019
Date

07 May 2019
Date

07 MAY 2019
Date

07 May 2019
Date

07 May 2019
Date

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Term	Definition/Explanation
ALP	alkaline phosphatase
ALT	alanine aminotransferase
ANOVA	analysis of variance
AST	aspartate aminotransferase
ATC	Anatomical Therapeutic Chemical
AUC	area under the curve
BMD	bone mineral density
BMI	body mass index
C_τ	predose trough concentrations
CDF	cumulative distribution function
CFI	comparative fit index
CI	confidence interval
CRF	case report form
CSR	clinical study report
CTCAE	common terminology criteria for adverse events
DSMB	data safety monitoring board
DXA	dual-energy x-ray absorptiometry
E2	estradiol
ECG	electrocardiogram
eCRF	electronic case report form
EDC	electronic data capture
eDiary	electronic diary
EOT	end-of-treatment
EQ-5D-5L	European Quality of Life Five-Domain Five-Level
FP	feminine product
FPRR	feminine product return rate
FSH	follicle-stimulating hormone
GFI	goodness of fit index
Hgb	hemoglobin
ICH	International Council on Harmonisation
ITT	intent-to-treat
KM	Kaplan Meier
LH	luteinizing hormone
LLN	lower limit of normal
LS	least squares
max	maximum
MBL	menstrual blood loss
min	minimum
mITT	modified intent to treat

Term	Definition/Explanation
MedDRA	Medical Dictionary for Regulatory Activities
mmHg	millimeters of mercury
M-vol	myoma volume
NET	norethindrone
NETA	norethindrone acetate
NRS	Numerical Rating Scale
PBO	placebo
PDF	probability density function
PGA	patient global assessment
PK	pharmacokinetic
PT	Preferred Term
QD	once daily
QTcF	corrected QT interval Fridericia
RMSEA	root mean square error of approximation
SAP	statistical analysis plan
SD	standard deviation
SES	standardized effect size
SMQ	standard MedDRA query
SOC	System Organ Class
UFS-QoL	Uterine Fibroid Symptom and Health-Related Quality of Life (Questionnaire)
ULN	upper limit of normal
U-vol	uterine volume
WHO	World Health Organization
Wks	weeks

1. INTRODUCTION

The purpose of this statistical analysis plan (SAP) is to describe the analyses planned for phase 3 studies MVT-601-3001 (LIBERTY 1) and MVT-601-3002 (LIBERTY 2), both entitled "An International Phase 3 Randomized, Double-Blind, Placebo-Controlled Efficacy and Safety Study to Evaluate Relugolix Co-Administered with and without Low Dose Estradiol and Norethindrone Acetate in Women with Heavy Menstrual Bleeding Associated with Uterine Fibroids." In these studies, patients are randomized to one of three treatment arms: relugolix 40 mg + estradiol/norethindrone acetate (E2/NETA) 1 mg/0.5 mg for 24 weeks (Group A, also referred to as the relugolix + E2/NETA group), relugolix 40 mg for 12 weeks followed by 12 weeks of relugolix 40 mg + E2/NETA 1 mg/0.5 mg (Group B, also referred to as the relugolix + delayed E2/NETA group), or placebo for 24 weeks (Group C, also referred to as the placebo group).

The 2 phase 3 studies are replicative; the only difference between the two protocols is the Week 24 endometrial biopsies, which in MVT-601-3001 are done in all patients and in MVT-601-3002 depend on the results of the Week 24 ultrasound.

This SAP was developed in accordance with the International Council on Harmonisation (ICH) E9 guidelines. All decisions regarding statistical analysis of the study, as defined in this SAP, will be made prior to unblinding of the study data.

The SAP is based on:

- Protocol MVT-601-3001, Amendment 2, dated 18 Sept 2017;
- Protocol MVT-601-3002, Amendment 2, dated 25 Sept 2017;
- ICH guidelines E3 (Clinical Study Reports) and E9 (Statistical Principles for Clinical Trials).

This document may evolve over time (eg, to reflect the requirements of protocol amendments or regulatory requests). However, the SAP is to be finalized, approved by the sponsor, and placed on file before the database is locked. Changes to the final approved plan will be noted in the clinical study report (CSR). Unless otherwise specified, the objectives, definitions of endpoints, and pre-specification of analyses presented in this document apply to both studies.

1.1. Study Objectives and Endpoints

The study objectives and corresponding endpoints are listed in the following table. The endpoints in *italics* are not listed in the protocol, but they have been identified as important for assessment of treatment effect on the basis of emerging data and clinical relevance to the study objectives and therefore are included in this SAP.

Table 1: Study Objectives and Endpoints

Objective(s)	Endpoint(s)
Primary Efficacy	
To determine the benefit of relugolix 40 mg once daily co-administered with E2 1 mg and NETA 0.5 mg compared with placebo for 24 weeks on heavy menstrual bleeding associated with uterine fibroids	Proportion of women in the relugolix + E2/NETA group versus the placebo group who achieve an MBL volume of < 80 mL AND at least a 50% reduction from Baseline MBL volume over the last 35 days of treatment, as measured by the alkaline hematin method
Key Secondary Efficacy (Alpha-Protected for Hierarchical Hypothesis Testing — relugolix + E2/NETA versus placebo)	
Achievement of amenorrhea	Proportion of women who achieve amenorrhea over the last 35 days of treatment
Heavy menstrual bleeding associated with uterine fibroids	Percent change from Baseline to Week 24 in MBL volume
Impact of uterine fibroids on symptoms, activities, and health-related quality of life as measured by components of the UFS-QoL	*Change from Baseline to Week 24 in the UFS-QoL Bleeding and Pelvic Discomfort Scale score, a sub-scale of the UFS-QoL Symptom Severity scale*
Change in hemoglobin	*Proportion of women with a hemoglobin ≤ 10.5 g/dL at Baseline who achieve an increase of > 2 g/dL from Baseline to Week 24*
Pain associated with uterine fibroids	*Proportion of patients with a maximum NRS score ≤ 1 during the last 35 days before the last dose of study drug in the subset of women with a maximum NRS score ≥ 4 for pain associated with uterine fibroids during the last 35 days prior to randomization*
Uterine fibroid volume	Percent change from Baseline to Week 24 in uterine fibroid volume
Uterine volume	Percent change from Baseline to Week 24 in uterine volume
Other Secondary Efficacy (Not for Hierarchical Hypothesis Testing) [a]	
To determine the benefit of relugolix 40 mg once daily for 12 weeks followed by 12 weeks of relugolix 40 mg once daily co-administered with E2 1 mg and NETA 0.5 mg compared with placebo for 24 weeks on heavy menstrual bleeding associated with uterine fibroids	Proportion of women in the relugolix + delayed E2/NETA group versus the placebo group who achieve an MBL volume of < 80 mL AND at least a 50% reduction from Baseline MBL volume over the last 35 days of treatment, as measured by the alkaline hematin method
Heavy menstrual bleeding associated with uterine fibroids	• *Percent change from Baseline in MBL volume by visit* • *Change from Baseline in MBL volume by visit*

Objective(s)	Endpoint(s)
	• Time to achieve an MBL volume of < 80 mL AND at least a 50% reduction from Baseline MBL volume as measured by the alkaline hematin method • *Proportion of women in the relugolix + E2/NETA group versus the placebo group who achieve an MBL volume of < 80 mL AND at least a 50% reduction from Baseline MBL volume by visit*
Achievement of amenorrhea	• *Sustained amenorrhea rate by visit* • *Time to achieving sustained amenorrhea* • *Time to achieving amenorrhea*
Change in hemoglobin	• Proportion of women with a hemoglobin below the lower limit of normal at Baseline who achieve an increase of ≥ 1 g/dL from Baseline at Week 24 • *Change from Baseline to Week 24 in hemoglobin for women with a hemoglobin ≤ 10.5g/dL at Baseline*
Impact of uterine fibroids on symptoms, activities and health-related quality of life as measured by components of the UFS-QoL	• Change from Baseline to Week 24 in the UFS-QoL Symptom Severity Scale score • Change from Baseline to Week 24 in the UFS-QoL Activities Scale score • *Change from Baseline to Week 24 in the UFS-QoL Revised Activities Scale score* • *Proportion of responders who achieved a meaningful increase of at least 20 points from Baseline to Week 24 in UFS-QoL Revised Activities Scale score* • *Proportion of responders who achieved a meaningful reduction of at least 20 points from Baseline to Week 24 in UFS-QoL Bleeding and Pelvic Discomfort Scale score* • Change from Baseline to Week 24 in the interference of uterine fibroids with physical activities based on UFS-QoL Question 11 • Change from Baseline to Week 24 in the interference of uterine fibroids with social activities based on UFS-QoL Question 20

Objective(s)	Endpoint(s)
	• Change from Baseline to Week 24 in embarrassment caused by uterine fibroids based on UFS-QoL Question 29
Patient global assessment for function and symptoms as measured by the PGA for function and symptoms	• Change in PGA for uterine fibroid related function from Baseline to Week 24 • Change in PGA for uterine fibroid symptoms from Baseline to Week 24 • *Proportion of patients achieving improvement from Baseline in PGA for uterine fibroid symptoms from Baseline to Week 24* • *Proportion of patients achieving improvement from Baseline in PGA for uterine fibroid related function from Baseline to Week 24*
Impact of heavy menstrual bleeding on social, leisure, and physical activities as measured by the Menorrhagia Impact Questionnaire	• Change from Baseline to Week 24 in the Menorrhagia Impact Questionnaire Score for physical activities • Change from Baseline to Week 24 in the Menorrhagia Impact Questionnaire Score for social and leisure activities
Pain associated with uterine fibroids[b]	Proportion of women who achieve a *maximum* NRS score for pain associated with uterine fibroids over the last 35 days of treatment that is at least a 30% reduction from Baseline in the subset of women with a maximum pain score ≥ 4 during the 35 days prior to randomization
Safety	
To determine the safety of 24 weeks of relugolix 40 mg once daily co-administered with either 12 or 24 weeks of E2 1 mg and NETA 0.5 mg in women with heavy menstrual bleeding associated with uterine fibroids compared with placebo for 24 weeks	Treatment-emergent adverse events, change in vital signs (including weight), clinical laboratory tests, and electrocardiograms
To determine the percent change from Baseline to Week 12 in bone mineral density at the lumbar spine (L1-L4) in the relugolix + E2/NETA group compared with the relugolix + delayed E2/NETA group in women with heavy menstrual bleeding associated with uterine fibroids	Percent change from Baseline to Week 12 in bone mineral density at the lumbar spine (L1-L4) in the relugolix + E2/NETA group compared with relugolix + delayed E2/NETA group as assessed by DXA

Objective(s)	Endpoint(s)
To determine the change in bone mineral density of women with heavy menstrual bleeding associated with uterine fibroids treated with 24 weeks of relugolix 40 mg once daily co-administered with either 12 or 24 weeks of E2 1 mg and NETA 0.5 mg compared with placebo for 24 weeks	Percent change from Baseline to Week 24 in bone mineral density at the lumbar spine (L1-L4), total hip, and femoral neck as assessed by DXA
To determine the incidence of vasomotor symptoms with relugolix 40 mg once daily co-administered with either 12 or 24 weeks of E2 1 mg and NETA 0.5 mg in women with heavy menstrual bleeding associated with uterine fibroids	Incidence of vasomotor symptoms
Pharmacokinetic and Pharmacodynamic	
To evaluate the pharmacokinetic and pharmacodynamic effects of 24 weeks of relugolix 40 mg once daily when co-administered with either 12 or 24 weeks of E2 1 mg and NETA 0.5 mg	• Predose trough concentrations (C_τ) of relugolix, and NET and Baseline-adjusted E2 concentration • Absolute and changes from Baseline to Week 24 in predose concentrations of LH, FSH, E2, and progesterone
Exploratory	
To determine the benefit of 24 weeks of relugolix 40 mg once daily co-administered with either 12 or 24 weeks of E2 1 mg and NETA 0.5 mg compared with placebo on patient-reported quality of life outcome measures (EQ-5D-5L)	Change from Baseline to Week 24 in the EQ-5D-5L Scale score

Abbreviations: DXA, dual energy x-ray absorptiometry; E2, estradiol; EQ-5D-5L, European Quality of Life Five-Domain Five-Level; FSH, follicle-stimulating hormone; LH, luteinizing hormone; MBL, menstrual blood loss; NET, norethindrone; NETA, norethindrone acetate; NRS, numerical rating scale; PGA, Patient Global Assessment; UFS-QoL, Uterine Fibroid Symptom and Health-Related Quality of Life.

[a] The secondary endpoints below will be assessed comparing the relugolix + E2/NETA group with the placebo group inferentially; the relugolix + E2/NETA group to the relugolix + delayed E2/NETA group and the relugolix + delayed E2/NETA group to the placebo group descriptively, unless otherwise specified.

[b] Changed from mean NRS score (in the protocol) to maximum NRS score. Since pain associated with uterine fibroids is mostly during menstrual days, mean NRS scores over the last 35 days is very low (< 1) for most patients, hence, not appropriate to define percent reduction from Baseline.

2. STUDY DESIGN

2.1. Summary of Study Design

The LIBERTY 1 and LIBERTY 2 studies are two replicate, randomized, double-blind, placebo-controlled phase 3 studies evaluating the efficacy and safety of relugolix 40 mg in combination with E2 1 mg/NETA 0.5 mg in women with heavy menstrual bleeding associated with uterine fibroids (MVT-601-3001, MVT-601-3002). Patients with heavy menstrual bleeding associated with uterine fibroids — as evidenced by a menstrual blood loss (MBL) volume of ≥ 80 mL per cycle for 2 cycles or ≥ 160 mL during one cycle, as measured by the alkaline hematin method during the screening period — who met other eligibility criteria were randomly assigned (1:1:1) to 1 of the 3 treatment arms:

- Group A (relugolix + E2/NETA): relugolix 40 mg once daily co-administered with E2 1 mg/NETA 0.5 mg for 24 weeks;
- Group B (relugolix + delayed E2/NETA): relugolix 40 mg once daily for 12 weeks followed by relugolix 40 mg once daily co-administered with E2 1 mg/NETA 0.5 mg for 12 weeks;
- Group C (placebo): placebo for 24 weeks

Randomization was stratified as follows:

- Geographic Region: North America versus Rest of World;
- Mean screening MBL volume using alkaline hematin method: < 225 mL versus ≥ 225 mL.

The primary endpoint for both trials is the proportion of women receiving relugolix + E2/NETA (Group A) versus placebo (Group C) who achieve BOTH a MBL volume of < 80 mL AND at least a 50% reduction from Baseline in MBL volume over the last 35 days of treatment, as measured by the alkaline hematin method.

This study includes a screening period (up to ~13 weeks), a randomized treatment period (24 weeks), and a safety follow-up period (~30 days). During the screening period, diagnoses of uterine fibroids are confirmed by centrally reviewed transvaginal ultrasound. Women with iron-deficient microcytic anemia and hemoglobin ≥ 8 g/dL and ≤ 10 g/dL during the screening period are treated with oral or parenteral iron replacement therapy. After randomization, patients begin double-blinded study drug treatment for 24 weeks.

Patients who complete LIBERTY 1 or LIBERTY 2, including those randomized to placebo, and who meet other eligibility criteria are offered the opportunity to enroll in a 28-week open-label extension study, in which all patients will receive relugolix 40 mg co-administered with E2 1 mg and NETA 0.5 mg. Patients who do not enroll into the extension study have a safety follow-up visit approximately 30 days after their last doses of study medication.

Additional safety follow-up may be performed after the safety follow-up visit. Data collected during the additional safety follow-up period will be summarized and reported in an addendum to the respective clinical study report. Patients who are not proceeding into the extension study and who have endometrial hyperplasia or endometrial cancer on the endometrial biopsy should be treated as per standard of care and additional follow-up should be evaluated and managed, as

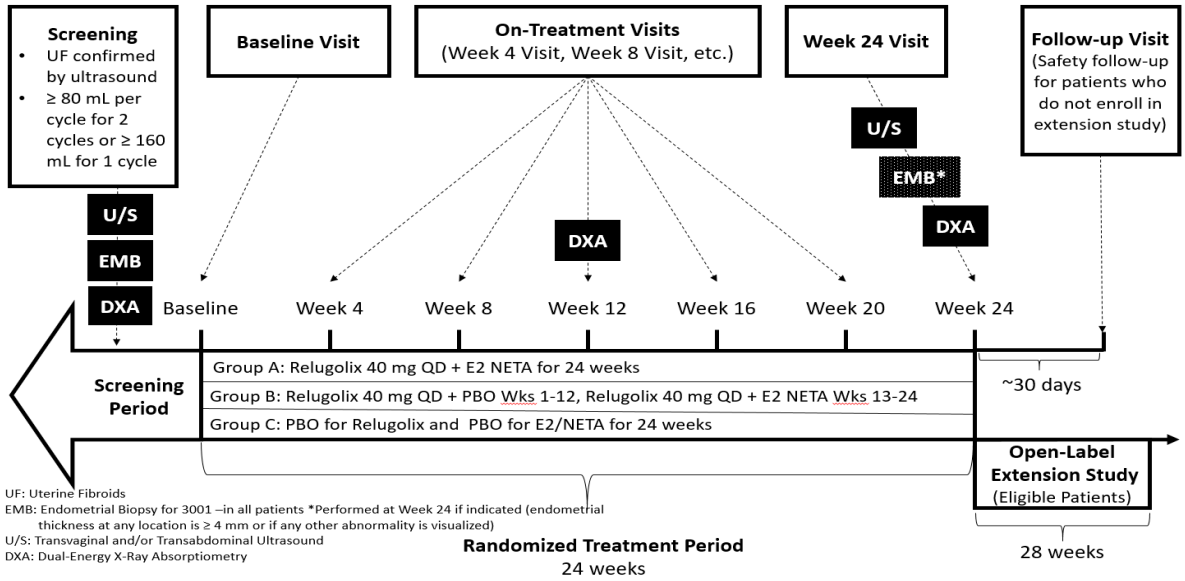
needed, by a gynecologist. In addition, they should undergo a repeat biopsy in 3 to 6 months after the Week 24/Early Termination and will be contacted to obtain information on procedures performed or treatments received (if any) for the biopsy findings through the time of the repeat biopsy. The repeat biopsy will be submitted to the central laboratory.

Patients who are not proceeding into the extension study and who have a bone mineral density (BMD) loss of > 2% at the lumbar spine (L1–L4) or total hip relative to the Baseline measurement at their Week 24/Early Termination visit will undergo a follow-up DXA scan 6 months (± 1 month) after discontinuation of study drug and will be contacted to obtain information about medications and conditions (eg, pregnancy, hyperparathyroidism, hypothyroidism, etc.) that might affect BMD through the time of the repeat DXA scan. If the DXA scan 6 months post-treatment continues to show BMD loss of > 1.5% at the lumbar spine and/or > 2.5% at the total hip compared with Baseline, patients will have an additional scan at 12 months post-treatment. All follow-up DXA scans will be submitted for central reading. Patients whose menses had not resumed as of the safety follow-up visit for unexplained reasons will be contacted by telephone to determine if menses have resumed. Patients with reductions in visual acuity will be referred for ophthalmology consultation.

An external independent data and safety monitoring board (DSMB) was established to review periodic safety analyses, including BMD assessments. The roles and responsibilities of the independent DSMB are described in a separate charter. A separate SAP was created to document the specific safety data analyses that would be performed by an independent data coordinating center for the DSMB on an ongoing basis during the study.

A schematic of the study is presented in Figure 1.

Figure 1: Study Schematic



Abbreviations: E2, estradiol; NETA, norethindrone; PBO, placebo; QD, once daily; Wks, weeks.

2.2. Sample Size Considerations

2.2.1. Sample Size Justifications for Primary Efficacy Endpoint

The following assumptions were used to determine the sample size for this study:

- 2-sided type I error rate: 0.05
- Randomization: 1:1:1
- Responder rate for placebo group: 25%
- Difference in responder rates between the relugolix + E2/NETA group and the placebo group: 30%
- Dropout rate: ~20%

With the assumption of a dropout rate of 20%, approximately 130 women in the relugolix + E2/NETA group and 130 women in the placebo group will provide at least 99% power at a 2-sided 0.05 significance level to detect a 30% difference in responder rates between relugolix + E2/NETA group and the placebo group for the primary endpoint. With an additional 130 women in the relugolix + delayed E2/NETA group, the total sample size will be approximately 390 women.

The assumed responder rate of 25% for the placebo group is within the range of responder rates observed from similar phase 3 trials in uterine fibroids (Stewart, 2017). The sample size and power calculations are based on a chi-squared test.

2.2.2. Sample Size Justifications for Percent Change in Bone Mineral Density at 12 Weeks

A pooled analysis of the percent change in BMD at 12 weeks using data from both phase 3 studies is described separately in the statistical analysis plan for the Integrated Summary of Safety. The results of this pooled analysis comparing the relugolix + E2/NETA group with the relugolix + delayed E2/NETA group will be presented in the Integrated Summary of Safety and will not be included in the CSRs for these studies.

For the comparison of the relugolix + E2/NETA group with the relugolix + delayed E2/NETA group with respect to the percent change in BMD from Baseline to Week 12 at the lumbar spine (L1–L4), approximately 260 women in the relugolix + E2/NETA group (pooled between the LIBERTY 1 and LIBERTY 2 studies) and 260 women in the relugolix + delayed E2/NETA (pooled) will provide at least 90% power at a 2-sided 0.05 significance level to detect a 1.25% absolute treatment difference, assuming a standard deviation of 4% and up to 15% dropout rate for each treatment group. Power calculations for this BMD comparison are based on a two-sample t-test.

Sample size and power calculations were performed using the software package *nQuery* 4.0 (Statistical Solutions Ltd.).

3. PLANNED ANALYSES

3.1. Interim Analyses

No interim efficacy analyses were planned or performed for these two studies.

An external, independent DSMB was established to review periodic safety analyses, including BMD assessments. A separate SAP was created to document the specific safety data analyses that would be performed by an independent data coordinating center for the DSMB on an ongoing basis during the study.

3.2. Final Analyses

The final analysis of all efficacy and safety data from MVT-601-3001 and MVT-601-3002 will occur after approximately 390 patients have been randomized to each study and have had the opportunity to be followed for 24 weeks of study treatment and through the 30-day safety follow-up visit. This document describes this final analysis.

There will be periodic safety data review by the DSMB. An independent data coordinating center has performed the periodic safety analyses and has provided results of these analyses to the DSMB, as defined in the DSMB charter and outlined in a separate DSMB SAP.

3.3. Safety Follow-Up Analyses

Patients who are not proceeding into the extension study and who have endometrial hyperplasia or endometrial cancer on the endometrial biopsy should be treated as per standard of care and additional follow up should be evaluated and managed, as needed, by a gynecologist. In addition, they should undergo a repeat biopsy in 3 to 6 months after the Week 24/Early Termination and will be contacted to obtain information on procedures performed or treatments received (if any) for the biopsy findings through the time of the repeat biopsy. The repeat biopsy will be submitted to the central laboratory.

Patients who are not proceeding into the extension study and who have a BMD loss of > 2% at the lumbar spine (L1–L4) or total hip relative to the Baseline measurement at their Week 24/Early Termination visit will undergo a follow-up DXA scan 6 months (± 1 month) after discontinuation of study drug and will be contacted to obtain information about medications and conditions (eg, pregnancy, hyperparathyroidism, hypothyroidism, etc) that might affect bone mineral density through the time of the repeat DXA scan. If the DXA scan 6 months post-treatment continues to show BMD loss of > 1.5% at the lumbar spine and/or > 2.5% at the total hip compared to Baseline patients will have an additional scan at 12 months post-treatment. All follow-up DXA scans will be submitted for central reading. Patients whose menses had not resumed as of the safety follow-up visit for unexplained reasons will be contacted by telephone to determine if menses have resumed. Patients with reductions in visual acuity will be referred for ophthalmology consultation.

Data collected during the additional safety follow-up period will be summarized and reported in an addendum to the respective clinical study report.

4. GENERAL CONSIDERATIONS FOR DATA ANALYSES AND HANDLING OF MISSING DATA

4.1. Data Presentation Conventions

All statistical analyses will be conducted using SAS® Version 9.2 or higher.

A statistical test for the primary and secondary efficacy endpoints will be assessed at a two-sided $\alpha = 0.05$ significance level, and all confidence intervals (CIs) will be reported as two-sided unless otherwise stated.

Where appropriate, variables will be summarized descriptively by study visit. For the categorical variables, the count and proportions of each possible value will be tabulated by treatment group. For continuous variables, the number of patients with non-missing values, mean, median, standard deviation (SD), minimum, and maximum values will be tabulated.

Unless otherwise specified, the following conventions will be applied to all analyses:

- Mean and median values will be formatted to one more decimal place than the measured value. Standard deviation values will be formatted to two more decimal places than the measured value; minimum and maximum values will be presented to the same number of decimal places as the measured value; if the measured value is large (eg, > 100), fewer decimal places may be displayed.
- Percentages will be rounded to 1 decimal place;
- p-values will be rounded to 4 decimal places. p-values < 0.0001 will be presented as "< 0.0001" and p-values > 0.9999 will be presented as "> 0.9999";
- 1 month = 30.4375 days. Month is calculated as (days/30.4375) rounded to 1 decimal place;
- 1 year = 365.25 days. Year is calculated as (days/365.25) rounded to 1 decimal place;
- Age will be calculated using the date of randomization. If only year of birth is collected, 1 July of the year of birth will be used to calculate age.
- 1 pound = 0.454 kg;
- 1 inch = 2.54 cm;
- Missing efficacy or safety data will not be imputed unless otherwise specified;
- For laboratory results above or below sensitivity limits displayed as "<" or ">" a quantification threshold, 0.0000000001 will be subtracted or added, respectively, to the threshold to derive a numeric result for analyses;
- For MBL volume reported as below the limit of quantification (for example, MBL below Quantification Level <5.0 mL or <2.5 mL), 0.0000000001 will be subtracted from the reported quantification threshold for the visit to derive a numeric result for analyses;
- For safety analyses, calculation of percentages will be calculated on the basis of the number of patients in the analysis population in each treatment group;

- For by-visit observed data analyses, calculation of percentages will be calculated on the basis of the number of patients with non-missing data as the denominator, unless otherwise specified;
- For other continuous endpoints, the summary statistics will include mean, SD, median, and range (minimum and maximum);
- For time-to-event endpoints, the summary statistics will include median time to event-free survival, 25th and 75th percentiles and number of patients at risk at specified time points;
- For categorical endpoints, the summary statistics will include counts and percentages;
- Confidence intervals, when presented, will generally be constructed at the 95% level. For binomial variables, exact methods will be employed, unless otherwise specified.

4.2. Analysis Populations

Three analysis populations are defined below. Number and percent of patients meeting the definition of each analysis population will be summarized by treatment group.

4.2.1. Modified Intent-to-Treat Population

Efficacy analyses will be performed using the modified Intent-to-Treat (mITT) population, unless otherwise specified. The mITT population is defined as all randomized patients who have received any amount of study drug (relugolix/placebo or E2/NETA/placebo). Efficacy analyses will be performed by treatment group as randomized.

4.2.2. Per-Protocol Population

The Per-Protocol population will consist of those members of the mITT population who do not have any of the specified subset of important protocol deviations (see Section 5.3).

The Per-Protocol population will not be analyzed if this population comprises > 95% or < 50% of the mITT population. The Per-Protocol population will be used for sensitivity analysis of the primary efficacy endpoint. The Per-Protocol population and the associated subset of important protocol deviations will be identified prior to unblinding the trial.

4.2.3. Safety Population

Safety analyses will be performed using the Safety population unless otherwise specified. The Safety population is the same as the mITT population and is defined as all randomized patients who have received any amount of study drug. Safety data will be analyzed by treatment group according to the actual treatment received (not the randomized treatment). Any patient who received at least one dose of relugolix will be considered as a relugolix patient.

4.3. Definitions, Computation, and Convention

4.3.1. Definition of Date of First Dose and Date of Last Dose of Study Drug

The date of the first dose of study drug is defined as the date when a patient receives the first dose of study drug (relugolix/placebo or E2/NETA/placebo). The date of the last dose of study

drug is defined as the date a patient receives the last dose of study drug. If the complete date of last dose of study drug is unknown, the last date the study drug was known to have been taken will be used.

4.3.2. Study Day

Study day will be calculated with respect to the date of the first dose of study drug (Study Day 1). For assessments conducted on or after the date of the first dose of study drug, study day will be calculated as:

(Assessment date – date of first dose of study drug) + 1

For assessments conducted before the date (and time) of the first dose of study drug, study day will be calculated as:

(Assessment date – date of first dose of study drug)

For patients who do not receive any amount of study drug, study day will be calculated as above with respect to the date of randomization.

4.3.3. Definition of Treatment Duration

Treatment duration is defined as the duration of time from the date of the first dose of study drug to the date of the last dose of study drug as follows:

(Date of last dose of study drug – Date of first dose of study drug) + 1

For patients without complete date of last dose of study drug, the last date study drug was known to have been taken will be used to calculate treatment duration. For patients who did not return for the Early Termination visits, the time after their last visit will not be included in calculations of treatment duration.

4.3.4. Definition of Baseline Value and Post-Baseline Value

Unless otherwise specified, Baseline values are defined as the last measurement before the first administration (date and time) of study drug. A post-Baseline value is defined as a measurement taken after the first administration of study drug. Change from Baseline is defined as (post-Baseline value – Baseline value). Both date and time of study drug administration and measurement will be considered when calculating Baseline value. If the time is not available, then the date alone will be used. For patients who receive no study medication, the date of randomization will be used in place of the date of first dose in determining Baseline and post-Baseline values.

4.3.5. Visit Windows

Visit windows, which will be used to associate assessments with a scheduled visit, will be used only for summarizing data by visit. The windows for scheduled assessments are shown in Table 2, Table 3 (electrocardiogram [ECG], BMD, Uterine Fibroid Symptom and Health-Related Quality of Life [UFS-QoL]), and Table 4 (transvaginal ultrasound, endometrial biopsy, and European Quality of Life Five-Domain Five-Level [EQ-5D-5L]), respectively. For both efficacy and safety assessments, the study day will be used to determine the associated visit window.

The data collected in the electronic diary (eDiary) related to bleeding and use of feminine products will be assigned to visit windows as specified in Table 5 and will be used to calculate the feminine product return rate (FPRR) as specified in Section 7.3.3.

If the results from more than one monthly or Week 12/Week 24 assessment are within a given visit window, the non-missing result from the assessment closest to the target date will be used. If two assessments are equally close to the target day, the earlier assessment will be used. For summaries of shift from Baseline in safety parameters, all values will be considered for these analyses.

Table 2: Visit Windows for Monthly Assessments

Visit	Start Day	Target Day	End Day
Week 4[a]	1	29	43
Week 8	44	57	71
Week 12	72	85	99
Week 16	100	113	127
Week 20	128	141	155
Week 24	156	169	196
Safety Follow-Up[b]	Date of last dose + 7 days	Date of last dose + 30 days	Date of last dose + 60 days

[a] Start day of Week 4 for study day 1 includes only post-Baseline assessments that occurred after the first dose.

[b] The safety follow-up visit window will be restricted to assessments prior to the date of initiation of another investigational agent or hormonal therapy affecting the hypothalamic-pituitary-gonadal axis or surgical intervention for uterine fibroids.

Table 3: Visit Windows for Week 12/Week 24 Assessments (ECG, BMD, UFS-QoL)

Visit	Start Day	Target Day	End Day
Week 12	64	85	106
Week 24	148	169	196
Safety Follow-up[a]	Date of last dose + 7 days	Date of last dose + 30 days	Date of last dose + 60 days

Abbreviations: BMD, bone mineral density; ECG, electrocardiogram; UFS-QoL, Uterine Fibroid Symptom and Health-Related Quality of Life.

[a] The safety follow-up visit window will be restricted to assessments prior to the date of initiation of another investigational agent or hormonal therapy affecting the hypothalamic-pituitary-gonadal axis or surgical intervention for uterine fibroids

Table 4: Visit Windows for Week 24 Assessments (Transvaginal Ultrasound, Endometrial Biopsy, EQ-5D-5L)

Visit	Start Day	Target Day	End Day
Week 24	128	169	196
Safety Follow-up[a]	Date of last dose + 7 days	Date of last dose + 30 days	Date of last dose + 60 days

Abbreviations: EQ-5D-5L, European Quality of Life Five-Domain Five-Level.

[a] The safety follow-up visit window will be restricted to assessments prior to the date of initiation of another investigational agent or hormonal therapy affecting the hypothalamic-pituitary-gonadal axis or surgical intervention for uterine fibroids.

Table 5: Time Window for eDiary and Feminine Product Collection

Visit	Feminine Product Collection Visit Date[a,b]	Time Window[a]
Week 4	X_1	(Date of Study Day 1) - < X_1
Week 8	X_2	(X_1 +1) - ≤ X_2
Week 12	X_3	(X_2 +1) - ≤ X_3
Week 16	X_4	(X_3 +1) - ≤ X_4
Week 20	X_5	(X_4 +1) - ≤ X_5
Week 24	X_6	(X_5 +1) - ≤ X_6
Week 24/EOT	X_{Last}[c]	(Previous Feminine Product Returned Visit +1)] – ≤ X_{Last}

[a] If feminine products are collected at more than 1 visit within a given visit window (Table 2), the last feminine product collection date will be used to define the time window. If the patient missed the previous visit, a planned study visit date will be used to calculate the window.

[b] In the absence of feminine product collection due to amenorrhea the visit date when amenorrhea was reported will be used.

[c] Date of last non-missing feminine product collection within the interval from (last dose date – 35) to (last dose date + 7 days) (see Section 7.3.3).

4.4. General Rules for Missing Data

Handling of missing data for the primary efficacy analysis is described in Section 7.3.5.

4.4.1. By-Visit Endpoints

By-visit endpoints will be analyzed using observed data, unless otherwise specified. For observed data analyses, missing data will not be imputed and only the observed records will be included.

4.4.2. Adverse Events and Concomitant Medications

The following imputation rules for the safety analyses will be used to address the issues with partial dates. The imputed dates will be used to determine the treatment-emergent period. For adverse events with a partial date, available date parts (year, month, and day) of the partial date will be compared with the corresponding date components of the start date and end dates of the treatment-emergent period to determine if the event is treatment emergent. When in doubt, the adverse event will be considered treatment emergent by default.

The following rules will be applied to impute partial dates for adverse events:

- If start date of an adverse event is partially missing, impute as follows:
  - If both Month and Day are missing and Year = Year of treatment start date, then set to treatment start date as long as adverse event end date is not prior to treatment start date;
  - If both Month and Day are missing and Year ≠ Year of treatment start date, then set to January 1;
  - If Day is missing and Month and Year = Month and Year of treatment start date, then set to treatment start date as long as adverse event end date is not prior to treatment start date;
  - If Day is missing and Month and Year ≠ Month and Year of treatment start date, then set to first of the month;
  - If start date is completely missing, set to treatment start date as long as adverse event end date is not prior to treatment start date.
- If end date of an adverse event is partially missing, impute as follows:
  - If both Month and Day are missing, then set to December 31;
  - If only Day is missing, then set to last day of the month;
  - If end date is completely missing, do not impute.

When the start date or end date of a medication is partially missing, the date will be imputed to determine whether the medication is prior or concomitant (or both).

The following rules will be applied to impute partial dates for medications:

- If start date of a medication is partially missing, impute as follows:
  - If both Month and Day are missing, then set to January 1;

  - If only Day is missing, then set to the first of the month.
- If end date of a medication is partially missing, impute as follows:
  - If both Month and Day are missing, then set to December 31;
  - If only Day is missing, then set to last day of the month.

If start date or end date of a medication is completely missing, do not impute.

5. STUDY POPULATION

5.1. Subjects Disposition

The number of patients for each of the following categories will be summarized by treatment group:

- All randomized patients;
- Patients included in the Safety population;
- Patients who completed the 12-Week randomized treatment period;
- Patients who completed the 24-Week randomized treatment period;
- Patients who discontinued early from the 24-Week randomized treatment period and reasons for discontinuation;
- Patients who enrolled in the extension study;
- Patients who entered the Post-Treatment Follow-Up Period and did not enroll in the extension study.

Patient disposition will be summarized for all randomized patients. Summaries will include the number and percentage of patients in the mITT and Safety populations. The number and percentage of patients who prematurely discontinue study drug and the reasons for discontinuation will be summarized by treatment group. The number and percentage of patients who continue into the extension study (MVT-601-3003) will also be summarized by treatment group.

5.2. Screen Failure

Reasons for screen failure will be summarized. Number and percentage of patients who did not pass screening will be based on the patients who signed the informed consent form but were not randomized.

5.3. Protocol Deviations

Protocol deviations will be categorized as important or minor per the protocol deviation plan. Important protocol deviations will include, but will not be limited to, the following categories:

- Randomized patient who did not satisfy key entry criteria;
- Randomized patient who met withdrawal criteria during the study but was not withdrawn;
- Randomized patient who received the wrong treatment;
- Randomized patient who received a prohibited concomitant medication that met criteria for an important protocol deviation;
- Unintentional unblinding of treatment assignment.

Important protocol deviations will be summarized by deviation category for all patients in the mITT population. A patient listing of all important protocol deviations will be provided.

In addition, patient eligibility, including inclusion criteria that are not met and exclusion criteria that are met at randomization enrollment, will be summarized for all patients in the mITT population.

A selected subset of the major protocol deviations that are likely to affect analysis of efficacy will be identified to define the Per-Protocol population prior to the database lock. This subset will include but will not be limited to the following important protocol deviations:

- Did not satisfy key entry criteria **(**restricted to patients with missing Baseline MBL volume or ineligible Baseline MBL volume);
- Drug compliance < 75%;
- Patient received prohibited concomitant medications that met criteria for important protocol deviation: restricted to patients who received prohibited concomitant medications that may cause significant drug-drug interaction;
- Unintentional unblinding of treatment assignment.

5.4. Demographic and Baseline Characteristics

Demographic and Baseline characteristics will be summarized by treatment group for the mITT population. Categorical data will be summarized using frequencies and percentages, by treatment group and overall (see Table 6 below). Summaries of continuous data will display the mean, SD, median, minimum, and maximum. The numbers of missing values will also be summarized.

Table 6: Categories for Demographic and Baseline Characteristics

Variable	Category
Age (years)	< 40, ≥ 40
Geographic region	North America, Rest of World
Race	Black or African American, White, American Indian or Alaska Native, Asian, Native Hawaiian or Other Pacific Islander, Other
Ethnicity	Hispanic or Latino, Not Hispanic or Latino or Not reported
BMI (kg/m^2) at Baseline	< 18.5, 18.5 to <25, 25 to <30, 30 to < 35, 35 to < 40, ≥ 40
History of prior pregnancy	Yes, No
Disease duration of uterine fibroid (years)	Min to <1, ≥ 1 to < 3, ≥3 to <5, ≥5 to <10, ≥ 10
Type of uterine fibroids	
Subserous fibroid	Yes, No
Intramural fibroid	Yes, No
Submucosal fibroid	Yes, No
Other	Yes, No
Any surgery for uterine fibroids	Yes, No
Volume of myoma at Baseline (cm^3)	< 25, ≥ 25
Volume of uterus at Baseline (cm^3)	< 300, ≥ 300
Menstrual blood loss volume at Baseline (mL)	< 225, ≥ 225
Menstrual blood loss volume at Baseline (mL)	< 160, ≥ 160
Hemoglobin at Baseline (g/dL)	Min to < 8, ≥ 8 to <10.5, ≥ 10.5 to <12, ≥ 12
UFS-QoL Bleeding and Pelvic Discomfort Scale	0 to < 25, 25 to <50, 50 to <75, 75 to 100
Maximum NRS score for uterine fibroid-associated pain at Baseline	< 4, ≥ 4
Patient Global Assessment	
Function	No limitation at all, mild limitation, moderate limitation, quite a bit of limitation, extreme limitation
Symptoms	Not severe, mildly severe, moderately severe, very severe, extremely severe

Abbreviations: BMI = body mass index; NRS = Numerical Rating Scale; UFS-QoL = Uterine Fibroid Symptom and Health-Related Quality of Life.

5.5. Medical History

Medical history will be coded using the Medical Dictionary for Regulatory Activities (MedDRA) and will be summarized by system organ class (SOC) and preferred term (PT). Additionally, summaries of uterine fibroid–specific medical and surgical treatment history will be provided. A patient with multiple occurrences of medical history within a PT will be counted only once in that PT.

5.6. Prior Medications and Concomitant Medications

Prior medications and concomitant medications taken during the study treatment period will be summarized for all patients in the Safety population by treatment group. Medications are considered concomitant if exposure occurs during the treatment period.

The number and percentage of patients who took at least one dose of a prior medication for treatment of uterine fibroids will be summarized by treatment group and overall using the World Health Organization (WHO) Drug Dictionary and summarized according to the Anatomical Therapeutic Chemical (ATC) Classification System and generic medication name. A patient who has been administered several medications with the same preferred medication name will be counted only once for that preferred medication name.

6. STUDY DRUG EXPOSURE AND COMPLIANCE

Patients in the Safety population will be summarized for extent of exposure and compliance to study drug by actual treatment received. Exposure to and compliance with relugolix (or relugolix placebo) and E2/NETA (or placebo) will be summarized separately and will be based on the drug accountability case report forms.

Study drug exposure summaries will include the total dosage taken in milligrams, the total number of tablets (or capsules) taken, and the treatment duration.

Study drug compliance will be summarized for the treatment period and will be calculated as follows:

(total tablets taken / total tablets expected to be taken) x 100

The total tablets taken will be calculated as:

(total tablets dispensed - total tablets returned)

The total tablets expected to be taken is calculated as the total number of tablets a patient is expected to take each day times the length of time (in days) that the patient was in the treatment period of the study. Tablets that were dispensed and not returned will be assumed to have been taken. For patients who did not return for their last scheduled visit, tablets that were dispensed and not returned will not be included in the calculation of study drug compliance. For patients who did not return for any post-Baseline visits and did not return dispensed study drug, study drug compliance will not be calculated and will be categorized as “not able to calculate” in summaries of study drug compliance.

Summary statistics of study drug compliance (eg, mean, median, etc.) will be presented, along with a categorical summary (eg, ≤ 80%, 80 to 100%, > 100%).

7. EFFICACY ANALYSES

7.1. General Considerations

Efficacy analyses will be conducted on the mITT population according to the randomized treatment assignment. Stratified analyses will incorporate the randomization stratification factors. If the group of patients at any factor level from a randomization stratification factor (eg, patients with Baseline MBL volume ≥ 225 mL) comprises < 10% of the entire mITT population, this stratification factor (eg, Baseline MBL volume) will not be used for stratified analyses. In addition, if there are < 15 patients in 1 of the 4 strata (derived from the 2 stratification factors each with 2 levels), only stratification factor of Baseline MBL volume (< 225 versus ≥225 mL) will be used in the stratified analysis for more robust strata-adjusted estimation of treatment effect. The stratification category used at the time of randomization (in the Interactive Web Recognition Service [IWRS] system) will be used for all analyses rather than data recorded on the electronic case report form (eCRF) unless otherwise specified. A sensitivity analysis of the primary endpoint will be performed if the data in the IWRS and eCRF for stratification factors differ by > 5%.

7.1.1. Analyses for Binary Data and Other Categorical Data

Binary data will be summarized by frequency counts and percentages for each treatment group.

7.1.2. Analyses for Categorical Data

Qualitative variables will be summarized by frequency counts and percentages. Unless otherwise specified, the calculation of proportions will include the missing category. Therefore, counts of missing observations will be included in the denominator and presented as a separate category.

7.1.3. Analyses for Continuous Data

Continuous variables will be summarized using descriptive statistics (eg, *n*, mean, median, SD, minimum, maximum, and first and third quartiles). For the analyses of change from Baseline, the mean at Baseline will be calculated for all patients with at least one post-Baseline value by treatment group. Additionally, the mean will also be calculated for each visit, including only the patients who are in the analysis who have data for that visit by treatment group.

7.1.4. Analyses for Time to Event Data

Time-to-event endpoints will be summarized using the Kaplan-Meier method. The median, quartiles, and probabilities of an event at particular time points will be estimated by the Kaplan-Meier method.

Confidence interval for the Kaplan-Meier estimation is calculated using the exponential Greenwood formula via log-log transformation of the survival function.

The variance of the treatment difference will be calculated using the following formula:

$$V\left[\widehat{S_R}(t) - \widehat{S_L}(t)\right] = \hat{V}\left[\widehat{S_R}(t)\right] + \hat{V}[\widehat{S_L}(t)];$$

where each of the component of the variance of the Kaplan-Meier estimate will be calculated using Greenwood's formula:

$$\hat{V}[\hat{S}(t)] = \hat{S}(t)^2 \sum_{t_i \le t} \frac{d_i}{n_i(n_i - d_i)}$$

where n_i denotes the number of patients at risk at time t_i, and d_i denotes the number of events observed at time t_i.

The 95% CI of the treatment difference will be calculated using a log-log transformation of the difference in survival function, as follows:

$$[(\widehat{S_R}(t) - \widehat{S_L}(t))^{exp(1.96\,\hat{\tau}(t))}, (\widehat{S_R}(t) - \widehat{S_L}(t))^{exp(-1.96\,\hat{\tau}(t))}]$$

$$\text{where } \hat{\tau}^2(t) = \frac{\hat{V}[\widehat{S_R}(t) - \widehat{S_L}(t)]}{\{[\widehat{S_R}(t) - \widehat{S_L}(t)]\log[\widehat{S_R}(t) - \widehat{S_L}(t)]\}^2}.$$

A stratified log-rank test will be used to compare each relugolix arm to placebo. Randomization stratification factors will be used to stratify inferential testing.

7.2. Multiplicity Adjustment

The primary and the ranked secondary efficacy analyses will be performed at an overall alpha level of 0.05 (two-sided) comparing relugolix + E2/NETA (Group A) with placebo (Group C). A test will be deemed statistically significant if the two-sided p-value rounded to four decimal places is < 0.05. A gate-keeping testing procedure will be applied to maintain the family-wise type I error rate for the testing of primary and ranked secondary endpoints (see Section 7.4.1 for details).

Comparative statistics (p-values, 95% CIs for differences) will be provided for the treatment comparison of relugolix + E2/NETA with placebo for all other secondary efficacy endpoints. A treatment comparison of relugolix + delayed E2/NETA (Group B) with placebo will be performed only for the primary efficacy endpoint. There will be no statistical testing for treatment differences between the relugolix groups (Group A versus Group B) for any efficacy endpoints. The relugolix + E2/NETA group and relugolix + delayed E2/NETA group will be compared for the following safety endpoints: percent change from Baseline to Week 12 in BMD and incidence of vasomotor symptoms by 12 weeks (see Section 9.3.5 and Section 9.1.7, respectively). The above comparative analyses are not part of the gate-keeping testing procedure for label claims. p-values for primary and key secondary endpoints were adjusted for multiplicity. All other p-values are provided at a nominal level of 0.05.

7.3. Primary Efficacy Endpoint

The primary efficacy endpoint of the study is the proportion of women who achieve an MBL volume of < 80 mL AND at least a 50% reduction from Baseline in MBL volume over the last 35 days of treatment as measured by the alkaline hematin method. The primary endpoint will be referred to as responder rate and derived on the basis of the total MBL volume measured at the Week 24/EOT visit window taking into consideration the patient's compliance with return of feminine products and completion of the eDiary (see Section 7.3.2 and Section 7.3.4 for details).

7.3.1. Primary Efficacy Analysis

The following primary hypothesis for the primary efficacy endpoint will be tested:

Null hypothesis H_{01}: $\pi_R \leq \pi_P$ versus Alternative hypothesis H_{a1}: $\pi_R > \pi_P$

where π_R and π_P are the responder rates at Week 24/EOT for relugolix + E2/NETA (Group A) and placebo (Group C), respectively.

The treatment comparison between the relugolix + E2/NETA and the placebo will be analyzed using a Cochran-Mantel-Haenszel test statistic for proportions stratified by the Baseline mean MBL volume using the alkaline hematin method (< 225 mL versus ≥ 225 mL) and geographic region (North America versus Rest of World). The difference in responder rates between the relugolix + E2/NETA and placebo and its two-sided 95% CI will be estimated using stratum-adjusted Mantel-Haenszel proportions. The unadjusted responder rates and the difference in responder rates between the relugolix + E2/NETA and placebo groups and the corresponding two-sided 95% CI also will be provided. The study will be considered positive if the treatment effect for the primary endpoint is statistically significant with two-sided p-value < 0.05.

For the primary analysis, primary endpoint will incorporate the missing data handling rules described in Section 7.3.5.

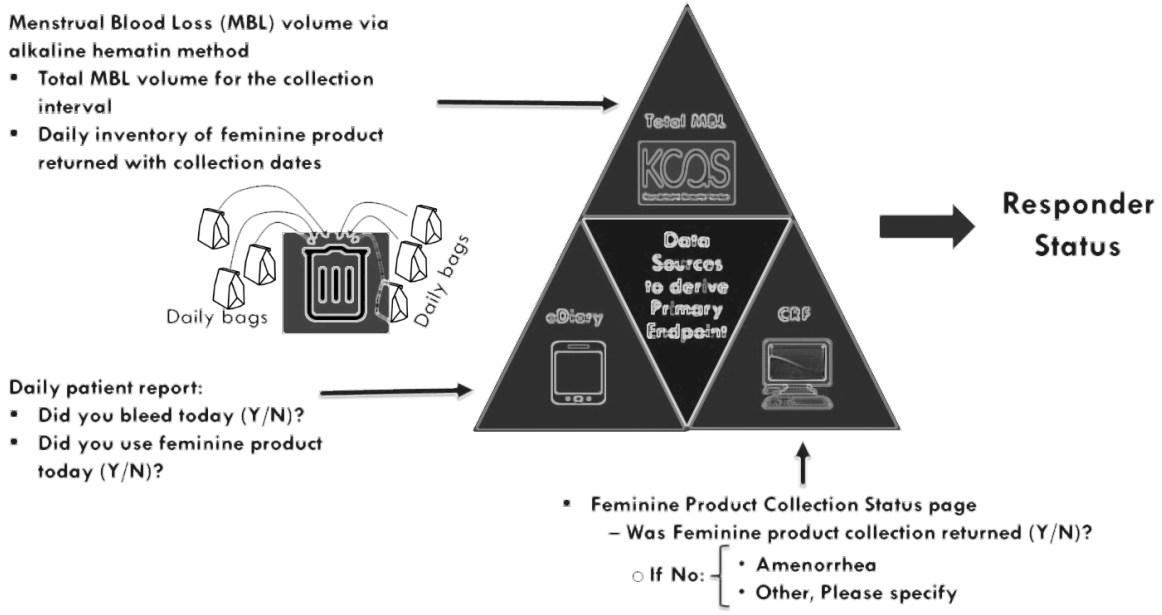
7.3.2. Data Sources Supporting Derivation of Responder Status

The data sources that will be used to support derivation of responder status, the primary endpoint, are depicted in Figure 2 below. They include:

- Menstrual blood loss volume determined by the alkaline hematin method;
- Daily patient report of bleeding (yes/no) and use of feminine product (yes/no) captured in the eDiary;
- The status of feminine product (FP) collection return (yes/no) recorded on the eCRF page at each visit with specific reasons captured when no product collection was returned.

The total MBL volume is reported from the analysis of FP returned for each collection interval. An inventory of days (with dates) for which FP was collected and returned is also available. This inventory is aligned with patients' reports of bleeding and FP use in the eDiary. The status of FP collection return, and specifically the reason for non-return of FP reported on the Feminine Product Collection eCFR page is used to support derivation of responder status (see Section 7.3.5 for details).

Figure 2: Data Sources Supporting Derivation of Primary Endpoint



Abbreviations: CRF = case report form.

7.3.3. Definitions Related to Menstrual Blood Loss

Menstrual Blood Loss Volume

All returned feminine products (validated, validated but unauthorized, or unvalidated products) collected at each visit will be analyzed by the alkaline hematin method to obtain the MBL volume. The MBL volume measured over the Week 24/EOT feminine product collection interval (up to 35 days prior to the last dose of treatment) will be used for analysis of the primary efficacy endpoint (see details below). The vendor, KCAS, reports when unauthorized feminine products (products not dispensed for use in the trial) have been returned. KCAS also reports whether the unauthorized products have previously been validated for their analysis. The report details MBL volumes for authorized, unauthorized but validated, and unauthorized and unvalidated products.

Validated Menstrual Blood Loss Volume

All returned feminine products collected at each visit, with the exception of unvalidated products, will be assessed by the alkaline hematin method to obtain the validated MBL volume. The validated MBL volume is derived from assessments of all returned validated feminine products (including validated and validated but unauthorized products) and will be used for sensitivity analysis.

Baseline Menstrual Blood Loss Volume

Baseline MBL volume is defined as the average MBL volume from the one or two consecutive screening menstrual cycles used to meet the inclusion criteria prior to the date of the first dose of study drug as assessed by the alkaline hematin method as follows:

For patients with MBL volume ≥ 160 mL during the screening period, the Baseline MBL volume is the last measurement collected before the first administration of study drug.

If the MBL volume is < 160 mL, the Baseline MBL volume is defined as the average of the MBL volume from the two screening menstrual cycles used to meet the inclusion criteria prior to the date of the first dose of study drug as assessed by the alkaline hematin method (see Figure 4-2 of the study protocol for details).

Week 24/EOT Feminine Product Collection Interval

To ensure collection of all feminine products used during that menstrual cycle, an interval of up to 35 days for measurement of the primary endpoint was selected to accommodate women who continue to have cyclic bleeding on study treatment and whose natural cycle was at the upper end of the normal cycle duration range. This method is consistent with that used during screening for collection of feminine products. Specifically, the feminine product collection interval at Week 24/EOT is driven by types of bleeding patterns experienced by the patients, as described below:

- For patients who continue to have cyclic bleeding, the length of the interval depends on the duration of the patient's natural cycle; this is consistent with the way the Baseline MBL volume was determined (eg, the interval ranging from approximately 21 to 35 days);
- Patients who report irregular, non-cyclic bleeding are instructed to collect and return all feminine product used between study visits, up to 35 days, as per the schedule of events;
- For patients who report amenorrhea on the feminine production collection eCRF page, an interval of last 35 days of treatment will be reviewed to ensure that reported amenorrhea is not due to incomplete collection.

For patients who are in the midst of an episode of cyclic bleeding at the time of the Week 24/EOT visit, the visit window may be extended up to 7 days after the last dose of study drug to ensure patients return all used feminine products over that bleeding episode.

Per protocol, all used feminine products are to be collected at each visit and returned for analysis using the alkaline hematin method. For patients who continue to have menstrual bleeding, study visits are timed such that the feminine products used in the entire menstrual bleeding cycle are collected in one container provided at each visit.

MBL Volume at Week 24/EOT

MBL volume at Week 24/EOT is defined as the MBL volume obtained from the feminine product returned over the Week 24/EOT feminine product collection interval, as described above. The MBL volume at Week 24/EOT will be used to derive the primary efficacy endpoint.

If a patient did not return feminine product over the last 35 days of treatment and reported amenorrhea on the feminine product return eCRF page, she will be considered as amenorrhoeic and her MBL volume will be assigned as 0 mL.

Feminine Product Return Rate at Week 24/EOT

To quantify degree of compliance with feminine product collection, the FPRR will be calculated based on the inventory of feminine product returned by day (dates) summarized on the Feminine Product Collection eCFR page (provided by the vendor, KCAS) and responses to the eDiary Question 4 regarding bleeding experience and Question 5 regarding the use of feminine product obtained for the corresponding eDiary window (see Table 5). Specifically:

- For those who returned feminine product at Week 24/EOT, the FPRR was calculated as the observed number of days with returned feminine products (based on the inventory of FP received by KCAS) divided by the expected number of days with bleeding and use of product as reported on the eDiary within the Week 24/EOT feminine product collection interval (as defined above).
- For those who did not return any feminine products:
  - If the reason was amenorrhea reported on the eCRF or if spotting/negligible bleeding was reported on the eCRF and confirmed by eDiary over the Week 24/EOT visit window, their FPRR will be set to 100% because the lack of menstruation obviates the need for feminine product collection.
  - Otherwise if the reason is any other, their FPRR was set to 0.

$$\text{FPRR} = \frac{\text{observed (No. of days with returned FP [per KCAS])}}{\text{expected (No. of days reported bleeding and use of FP [per eDiary])}} x100$$

Return of feminine products will be summarized in the CSR for Week 24/EOT visit.

7.3.4. Definition of Responder at Week 24/EOT

A responder at Week 24/EOT is defined as a patient who satisfies both the following:

- Had MBL volume of < 80 mL at Week 24/EOT;
- Had at least a 50% reduction from Baseline in MBL volume at Week 24/EOT.

The reduction from Baseline in MBL volume at Week 24/EOT will be calculated as the absolute change at Week 24/EOT in MBL volume from the Baseline MBL volume divided by the Baseline MBL volume.

Responder status at Week 24/EOT will be assessed based on the reported MBL volume at Week 24/EOT, in conjunction with treatment duration, compliance with feminine product collection, and compliance with eDiary entry over the same visit window (see Section 7.3.5 for details).

7.3.5. Derivation of Responder Status at Week 24/EOT and Missing Data Handling Rules

For the evaluation of primary endpoint, missing data handling rules will be implemented for deriving responder status at Week 24/EOT as described below. The following elements will be checked: duration of treatment exposure; compliance with feminine product collection against the eDiary, as measured by FPRR; compliance with eDiary entry, defined as the proportion of eDiary entry days over the length (days) of FP collection interval for Week 24/EOT visit; and reasons for no FP collection (as displayed in Table 7).

Patients with < 4 weeks of treatment who withdraw from the study prematurely due to lack of efficacy or withdraw from the study prematurely to undergo surgical intervention for uterine fibroids will be considered as non-responders.

All other patients will have their responder status determined as follows:

- For patients with a FPRR of 100%, responder status will be determined based on the observed MBL volume;
- For patients who had incomplete feminine product collection, with a FPRR of ≤ 100%, responder status will be derived based on either imputed or observed MBL volume;
  - Those with an MBL volume ≥ 80 mL or < 50% reduction from Baseline will be considered as non-responders;
  - Those with an MBL volume < 80 mL and ≥ 50% reduction from Baseline will be imputed for partial or complete missing MBL volume (see Section 7.3.6 for details).
- For patients who did not return a feminine product collection, responder status will be determined depending on the reason reported on the Feminine Product Collection eCRF:
  - If the reason is reported as Amenorrhea, the last 35 days of treatment will be used to derive responder status:
    - If the Week 24/EOT interval was 35 days, then she will be considered as a responder;
    - If the Week 24/EOT interval was <35 days, the following supportive information will be used to derive responder status:
      - If a patient reported amenorrhea at the visit prior to Week 24/EOT, she will be defined as a responder;
      - If a patient did not report amenorrhea at the visit prior to Week 24/EOT, eDiary data from the prior visit interval will be reviewed to confirm whether the patient was amenorrheic for a total of 35 days.
        - If the eDiary from the previous interval confirms amenorrhea, then the patient will be considered as a responder;

- Otherwise, MBL volume will be imputed.

- If the reason is Other and the specification describes spotting or negligible bleeding, responder status will be defined as follows:
    - The patient will be considered as a responder if it is supported by the eDiary data: the eDiary entry rate must exceed 70% and the patient must have reported no more than 5 total days of bleeding with product use and no more than 3 consecutive bleeding with product use over the collection interval.
    - If the eDiary entries did not confirm spotting or negligible bleeding, but the patient had at least 8 weeks of MBL volume data prior to the Week 24/EOT visit, her missing MBL volume will be imputed to determine responder status. Eight weeks of MBL volume data represents a reasonable minimum length of observation to justify imputation of the remaining data in assessing the effects of hormonal therapy.
    - Otherwise if the patient had < 8 weeks of MBL volume data, she will be considered as a non-responder;
- If the reason is any Other, the responder status will be derived as follows:
    - If the patient had at least 8 weeks of MBL volume data prior to the Week 24/EOT visit, her missing MBL volume will be imputed and her responder status will be based on the imputed MBL volume.
    - If the patient had < 8 weeks of MBL volume data, she will be considered as a non-responder.

Table 7: Derivation of Responder Status at Week 24/EOT and Missing Data Handling Rules – for Primary Analysis

Treatment Exposure	FP Collection (FPRR)	Observed MBL Volume	Reason for No FP Collection	Responder Status
< 4 weeks	N/A	N/A	N/A	Imputed as non-responder
≥ 4 weeks	100% FP Compliance	N/A	N/A	Based on the observed MBL volume
	<100% FP Compliance	MBL volume ≥ 80 mL or <50% reduction from Baseline	N/A	Imputed as non-responder based on the observed MBL volume
		MBL volume < 80 mL and ≥ 50% reduction from Baseline	N/A	**Based on the imputed MBL volume**
	No FP Collection	N/A	Reported "Amenorrhea"	Imputed as responder
			Reported "Spotting or negligible bleeding" and confirmed by eDiary[a]	Imputed as responder
			Reported "Spotting or negligible bleeding" although not confirmed by eDiary or any other reason, had at least 8 weeks of MBL volume data	**Based on the imputed MBL volume**
			The entries in the eDiary did not verify "Spotting or negligible bleeding" or any other reason and if had < 8 weeks of MBL volume data	Imputed as non-responders

Abbreviations: eDiary, electronic diary; FP, feminine product; EOT, end of treatment; MBL, menstrual blood loss; N/A, not available.

[a] Defined as those patients who meet the following criteria: eDiary entry rate > 70% and no more than 3 consecutive days and no more than 5 days of bleeding/spotting and use of feminine product reported on the eDiary over the Week 24/EOT visit window (see Table 5).

7.3.6. Mixed-Effects Model for Imputing Missing or Partially Missing MBL Volume at Week 24/EOT

For the primary analysis, patients with missing MBL volumes at Week 24/EOT will be identified per missing data handling rules as described above. For imputing missing data for the primary analysis, a mixed-effects model approach will be used, as the mixed-effects approach may better describe the effects of a hormonal treatment (such as suppression of the hypothalamic-pituitary-ovarian axis by GnRH antagonists).

Specifically, a mixed-effects model with repeated measures of MBL volumes at multiple time points (Weeks 4, 8, 12, 16, 20 and 24) will be fitted to predict percent change in MBL volume from Baseline (as a dependent variable) through the fixed-effects associated with covariates (ie, stratification factors of Baseline MBL volume and geographic region, visit, treatment, and visit by treatment interaction) and random effects (from the individual patients). In this model, an unstructured variance-covariance matrix is assumed for each patient.

See sample SAS codes below for illustration where PCHG_MBL is percent change in MBL volume from Baseline as a dependent variable, PID is patient identification number, BMBL is a randomization stratification factor (Baseline MBL < 225 vs ≥ 225), REGION is a randomization stratification factor (North America vs Rest of World), TRT is treatment group (relugolix + E2/NETA or Placebo), VISIT is visit time point (4, 8, 12, 16, 20, and 24 weeks) and TRT*VISIT is the visit by treatment interaction. The specification of type=UN implements unstructured variance-covariance matrix for an individual patient with multiple measures of MBL volumes.

```
proc mixed data=MBL_dataset method=REML covtest;
class PID BMBL REGION TRT VISIT;
model PCHG_MBL= BMBL REGION VISIT TRT VISIT*TRT/s outp=ufmi_mixed_p
covb;
repeated VISIT /type=UN subject=PID r;
lsmeans TRT/diff;
ods output SolutionF=mixparms CovB=mixcovb;
```

Applying this model over the observed longitudinal MBL volume data, the fixed-effects will be estimated and relationship of percent change in MBL volume from Baseline with the covariates will be characterized by the fitted model. From the fitted model, the percent change in MBL volume (whether missing or not) will be predicted for each patient at each visit and in a particular stratum. The imputed MBL volume will be obtained by first multiplying the imputed percent change with the individual patient's Baseline MBL volume to the difference, and then adding the Baseline BML volume to the difference.

The main reason for using percent change in MBL volume over reported MBL volume as a dependent variable in the mixed-effects model is that the percent change is part of the derivation of the primary endpoint. Secondly, the percent change is a normalized value adjusted for the Baseline value and less influenced by Baseline MBL volume, and therefore it is a better metric to describe the relationship of MBL volume reduction with hormonal treatment and to impute the missing volumes in a more robust fashion.

Since the purpose of using a mixed-effects model is imputing the missing MBL volumes identified at Week 24/EOT, the predicted MBL volumes at the corresponding Week 24/EOT visit will be used to determine responder status. For patients without the need for imputation, their responder status will be derived according to the algorithms laid out in Table 7. This imputation approach is consistent with the definition of responder at Week 24/EOT for the primary analysis.

7.3.7. Sensitivity Analyses

To assess the robustness of the primary analysis, the following sensitivity analyses of the primary endpoint will be conducted at Week 24/EOT.

7.3.7.1. Sensitivity Analysis 1

To assess the potential impact of unvalidated feminine product use, the primary endpoint will be analyzed as sensitivity analysis in a similar fashion to the primary analysis using the Week 24/EOT validated MBL volume (obtained from the validated or validated-but-unauthorized feminine products only and excluding unvalidated products).

7.3.7.2. Sensitivity Analysis 2

To assess the potential impact of missing data due to inadequate collection of feminine products, the primary endpoint will be analyzed with a sensitivity analysis using the missing data handling rules as described in Table 8 below where the observed MBL volume will be used to assess the responder status at Week 24/EOT when feminine product collection was incomplete. These rules differ from those used in the primary analysis in that no imputation will be implemented for patients with < 100% feminine product compliance and the reported MBL volume both < 80 mL and a ≥ 50% reduction from Baseline as highlighted in Table 8.

Table 8: Derivation of Responder Status at Week 24/EOT and Missing Data Handling Rules – for Sensitivity Analysis

Treatment Exposure	FP Collection (FPRR)	Observed MBL Volume	Reason for No FP Collection	Responder Status
< 4 weeks	N/A	N/A	N/A	Imputed as non-responder
≥ 4 weeks	100% FP Compliance	N/A	N/A	Based on the observed MBL volume
	< 100% FP Compliance	MBL volume ≥ 80mL or < 50% reduction from Baseline	N/A	Imputed as non-responder based on the observed MBL volume
		MBL volume < 80mL and ≥ 50% reduction from Baseline	N/A	**Based on the observed MBL volume**
	No FP Collection	N/A	Reported "Amenorrhea"	Imputed as responder
			Reported "Spotting or negligible bleeding" and confirmed by eDiary[a]	Imputed as responder
			Reported "Spotting or negligible bleeding" although not confirmed by eDiary or any other reason, had at least 8 weeks of MBL volume data	Based on the imputed MBL volume
			The entries in the eDiary did not verify "Spotting or negligible bleeding" or any other reason and if had < 8 weeks of MBL volume data	Imputed as non-responders

Abbreviations: eDiary, electronic diary; FP, feminine product; EOT, end of treatment; MBL, menstrual blood loss; N/A, not available.

[a] Defined as those patients who meet the following criteria: eDiary entry rate >70% and no more than 3 consecutive days and no more than 5 days of bleeding/spotting and use of feminine product reported on the eDiary over the Week 24/EOT visit window (see Table 5).

7.3.7.3. Sensitivity Analysis 3

To assess the potential impact of early discontinuation on the primary endpoint, the primary endpoint will be analyzed with a sensitivity analysis defining the patients' responder status as follows:

- Patients who discontinued study drug during the first 4 weeks for any reason or who discontinued study drug between Week 4 and Week 12 due to an adverse event, surgery or other intervention for heavy menstrual bleeding, reported lack of efficacy, or bleeding complaints will be considered as non-responders;
- All other patients will have their responder status defined using data from the Week 24/EOT assessment period using the last observation carried forward method.

7.3.7.4. Sensitivity Analysis 4

To assess the potential impact of the length and full exposure of the treatment, the primary endpoint will be analyzed for the Completers population as a sensitivity analysis. The Completers population is defined as patients in the mITT population who completed 24 weeks of study treatment.

7.3.7.5. Sensitivity Analysis 5

The primary endpoint will be analyzed on the Per-Protocol population as a sensitivity analysis, using the methods specified for the primary analysis (see definition of Per-Protocol population in Section 4.2.2).

7.3.7.6. Sensitivity Analysis 6

As a sensitivity analysis to the primary analysis using the mixed-effects model for imputing missing MBL volumes at Week 24/EOT, multiple imputation approach will be implemented as described below.

A multiple imputation method (Rubin, 1987; von Hippel, 2018) will be used to impute missing or partially missing MBL volume identified by the missing data handling rules (see Table 7 and Table 8) at Week 24/EOT as described in the following 5 steps. In this method, an arbitrary missing pattern will be assumed using Markov Chain Monte Carlo imputation to generate a monotone missing pattern for the observed longitudinal MBL volume values (including 0 mL if the patient has amenorrhea). Imputation will be performed separately by randomized treatment group (Sullivan, 2018), given the distinct bleeding patterns among the three treatment groups.

Normalizing transformations will be applied to the statistics estimated from each imputed dataset before the Rubin's combination rules can be applied (Ratitch, 2013). This combined estimation and statistical test will account for the additional variability due to imputation to provide a robust assessment of the treatment effect.

Step 1: Identifying patients with missing or incomplete MBL volume from the longitudinal MBL volume dataset as collected.

Step 2: Generating a monotone missing pattern using the Markov Chain Monte Carlo technique by imputing missing MBL volume measurements that are between non-missing results.

Step 3: Imputing the remaining missing values m = 100 times using a regression model; therefore, generating 100 complete longitudinal MBL volume datasets.

Note: if a patient missed Week 8 and prematurely discontinued study drug (eg, at Week 20) and MBL volume at Week 20 is missing or partially missing, MBL volume will be imputed for intermittent missing data at Week 8, Week 20 (EOT), and Week 24 due to discontinuation.

Step 4: Performing the same CMH test pre-specified for the primary endpoint analysis and estimating the responder rates for each arm using each of the 100 datasets based upon the MBL volume at Week 24/EOT.

Note: in the example above, the imputed MBL volume at Week 20 (EOT) will be used in the analysis, although MBL volume is imputed at Week 24.

Step 5: Combining the results from the 100 complete datasets to make inferences about the treatment effect on the responder rate.

7.3.8. Subgroup Analyses

Subgroup analyses of the primary efficacy endpoint comparing the relugolix + E2/NETA group versus the placebo group will be performed to assess whether treatment effects are consistent across clinically important subgroups. The odds ratio and its 95% CI based on a logistic regression model will be displayed in a forest plot for each subgroup. The logistic regression model will include treatment group, Baseline MBL volume value and geographic region as covariates. Subgroups will include, but will not be limited to, the subgroups outlined in Table 9.

Table 9: Planned Subgroup Analyses

Subgroup Name	Subgroup Level
Geographic region	North America vs Rest of World
Menstrual blood loss volume at Baseline (mL)	< 225 vs ≥ 225 < 120, 120 to < 160, 160 to < 225, ≥ 225
Age category (years)	< 40 vs ≥ 40 < 35, 35 to < 40, 40 to < 45, ≥ 45
Race	Black or African American vs Not Black or African American; Black or African American, White, Other
Volume of myoma at Baseline (cm^3)	< 25 vs ≥ 25
Volume of uterus at Baseline (cm^3)	< 300 vs ≥ 300
BMI (kg/m^2) at Baseline	< 30 vs ≥ 30 < 25, 25 to < 30, 30 to < 35, 35 to <40, ≥ 40
Maximum NRS score for uterine fibroid–associated pain at Baseline	< 4 vs ≥ 4
History of prior pregnancy	Yes/No

Abbreviations: BMI = body mass index; NRS = Numerical Rating Scale.

7.4. Secondary Efficacy Endpoints

Secondary efficacy variables include seven key secondary endpoints with alpha-protection and other secondary endpoints. All secondary efficacy endpoints and analyses are summarized in Appendix 1.

The treatment effect of relugolix + E2/NETA (Group A) compared to placebo (Group C) will be tested for the alpha-protected secondary endpoints using a gate-keeping procedure (see Section 7.4.1).

Comparative statistics (p-values, 95% CIs for differences) will be provided for treatment comparison of the relugolix + E2/NETA group with the placebo group for all other secondary efficacy endpoints. Treatment difference between the relugolix + delayed E2/NETA group and the placebo group will be formally tested only for the primary efficacy endpoint. There will be no statistical testing for treatment differences between the relugolix groups (relugolix + E2/NETA group against relugolix + delayed E2/NETA group) for any efficacy endpoint (see Section 7.4.2).

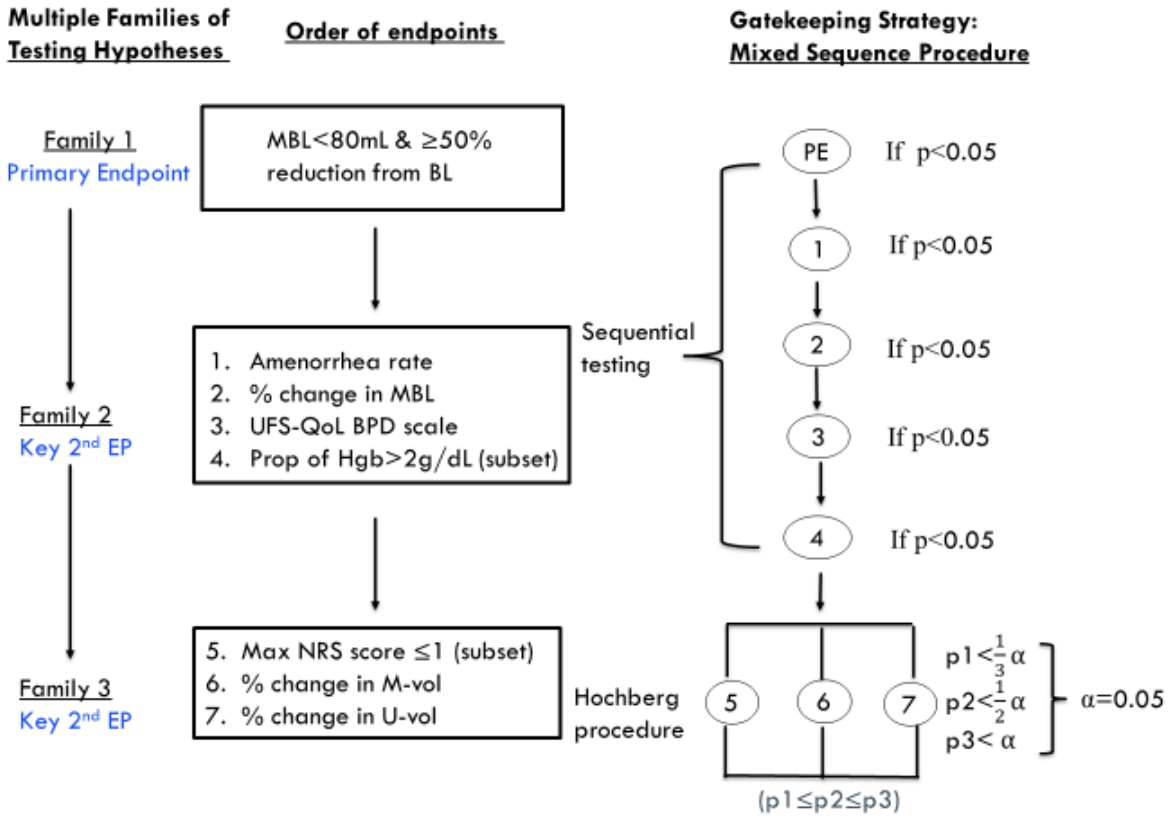
7.4.1. Key Secondary Efficacy Endpoints with Alpha-Protection

For testing whether relugolix + E2/NETA (Group A) is statistically significantly superior to placebo (Group C) for the primary efficacy endpoint as well as the seven key secondary endpoints listed below, a gate-keeping mixed sequence testing procedure will be applied to maintain the family-wise type I error rate. Under this testing procedure, the primary endpoint

will be tested first at a 2-sided 0.05 significance level. If the p-value for primary endpoint is < 0.05, the seven key endpoints listed below will be tested sequentially in the order depicted in Figure 3.

For the relugolix + E2/NETA group to be considered statistically superior to the placebo group on a secondary endpoint, the two-sided p-value must be < 0.05 for that secondary endpoint and for all higher-ranking secondary endpoints, as well as for the primary endpoint. If the two-sided p-value is < 0.05 for the fourth endpoint (proportion of women with a hemoglobin ≤ 10.5 g/dL at Baseline who achieve an increase of > 2 g/dL from Baseline at Week 24), the remaining three endpoints (the fifth, sixth, or seventh) will be tested using the Hochberg step-up procedure.

Figure 3: Mixed Sequence Testing Procedure for Primary and Key Secondary Endpoints



Abbreviations: BPD = Bleeding and Pelvic Discomfort; EP = endpoint; Hgb = hemoglobin; max = maximum; MBL = menstrual blood loss; M-vol = myoma volume; NRS = Numerical Rating Scale; PE = primary endpoint; Prop = proportion; UFS-QoL BPD = Uterine Fibroid Symptom and Health-Related Quality of Life Bleeding and Pelvic Discomfort; U-vol = uterine volume.

From the Hochberg procedure, the p-values will be calculated for the three endpoints (5, 6, and 7) and ranked from the smallest to the largest. The endpoint corresponding to the largest p-value gets tested first. If the p-value is < 0.05, then no further testing will occur, and it will be concluded that all three endpoints are positive. Otherwise, the endpoint corresponding to the second largest p-value will be tested. If the p-value is < 0.025, then no further testing will occur, and it will be concluded that the endpoints corresponding to the middle and smallest p-values are positive. Otherwise, the endpoint with the smallest p-value will be tested. If the p-value is < 0.0167, no further testing will occur, and it will be concluded that only the endpoint with the smallest p-value is positive. Otherwise, all three endpoints did not pass the statistical significance criterion at 0.05 level.

The seven key secondary efficacy endpoints are as follows:

1. Proportion of women who achieve amenorrhea over the last 35 days of treatment;
2. Percent change from Baseline to Week 24 in MBL volume;
3. *Change from Baseline to Week 24 in Bleeding and Pelvic Discomfort Scale score as measured by the UFS-QoL Symptom Severity Scale (Q1, Q2, Q5);*
4. *Proportion of women with a hemoglobin ≤ 10.5 g/dL at Baseline who achieve an increase of > 2 g/dL from Baseline at Week 24*
5. *Proportion of women who achieve a maximum NRS score ≤ 1 for uterine fibroid associated pain over the last 35 days of treatment in the subset of women with a maximum pain score ≥ 4 during the 35 days prior to randomization;*
6. Percent change from Baseline to Week 24 in uterine fibroid volume;
7. Percent change from Baseline to Week 24 in uterine volume.

For key secondary efficacy endpoints (1, 4, and 5) that are evaluating proportions, treatment comparisons will be performed using a stratified Cochran-Mantel-Haenszel test with the randomization stratification factors as strata. Point estimates and 2-sided 95% CIs for treatment differences in proportions will be provided.

For key secondary endpoint 4, an increase in hemoglobin of 2g/dL is considered clinically meaningful, because it corresponds to approximately the same increase as that expected after a transfusion of ~ 2 units of packed red blood cells (Man, 2016; Bachowski, 2017).

For deriving the key secondary endpoint 5 (proportion of women who achieve a maximum NRS score ≤ 1 for uterine fibroid-associated pain over the last 35 days of treatment in the subset of women with a maximum pain score ≥ 4 during the 35 days prior to randomization), the patient subset and Week 24/EOT maximum value are determined as follows.

Because patients were asked to begin eDiary entries after returning the first collection of feminine products, the number of eDiary entries made during screening varies with the duration of screening for each patient. Some patients required only one collection to be randomized, whereas others required as many as four collections to confirm eligibility.

Once the qualifying menstruation was completed and the patient qualified for randomization based upon resulting MBL volume(s), the recording of patient's NRS scores for screening phase

will be ended and the number of pain score days at Baseline can be as short as 7 days or as long as 70 days prior to randomization. If a patient meets the subset definition (maximum NRS score ≥ 4 at Baseline) over a portion of the screening days (eg, 7-70 days), she will also meet the subset definition on the entire 35 days interval.

Since the maximum NRS value is used to determine inclusion into the subset rather than an average NRS value, the variable number of days for inclusion of patients has no major impact on determining patient subset. To ensure robust estimate of response, the minimum number of non-missing daily pain scores required to calculate the maximum score at Week 24/EOT is at least 28 days (80% of the last 35 days of treatment) of pain scores recorded in the e-Diary entry.

The primary analysis of key secondary endpoint 5 will be analyzed for the subset of women who have a maximum pain score ≥ 4 during the 35 days prior to randomization and who have at least 28 days (80% of the last 35 days of treatment) of pain scores recorded in the e-Diary at Week 24/EOT. In addition, a sensitivity analysis will be conducted on the subset of women who have a maximum pain score ≥ 4 during the 35 days prior to randomization without restricting number of days of pain scores recorded in the e-Diary.

The analysis for endpoint 5 (proportion of women who achieve a maximum NRS score ≤ 1 for uterine fibroid-associated pain over the last 35 days of treatment in the subset of women with a maximum pain score ≥ 4 during the 35 days prior to randomization) will also be performed using NRS scores reported on eDiary during menstrual and non-menstrual days.

For key secondary efficacy endpoints (6 and 7) evaluating percent change from Baseline in uterine fibroid volume and uterine volume that are measured only at Week 24, an analysis of covariance (ANCOVA) model will be used to assess treatment effect with treatment, randomization stratification factors and Baseline value as covariates.

For key secondary efficacy endpoints (2 and 3) evaluating the change (absolute or % change) from Baseline to Week 24, treatment comparisons will be performed using a mixed model repeated measures approach with treatment, visit, randomization stratification factors and treatment by visit interactions included as fixed effects and random effects (from the individual patients). In this model, an unstructured variance-covariance matrix is assumed for each patient.

7.4.2. Other Secondary Efficacy and Exploratory Endpoints

The following describes the analysis methods for other secondary efficacy endpoints and exploratory endpoints. There are three types of analyses corresponding to the three types of endpoints (time-to-event, continuous and binary) (see Appendix 1 for details).

Time-to-Event Endpoint

For time to achieving an MBL volume of < 80 mL AND at least a 50% reduction from Baseline MBL volume (as an event), time to event will be defined as weeks from date of first dose of study drug to response (event) based on the MBL volume as assessed by the alkaline hematin method. The missing data handling rules described in Section 7.3.5 for deriving responder status at Week 24/EOT will be applied similarly at Weeks 8, 12, 16, and 20. Patients without an event will be censored at the last assessment date prior to the last dose of the study drug. Kaplan-Meier methods will be used to describe the time to event distributions. A log-rank test stratified by the randomization stratification factors using the proportional hazard model (p-value

from score test) will be used to compare relugolix + E2/NETA to placebo. Randomization stratification factors will be used to stratify inferential testing.

Continuous Endpoints

For endpoints evaluating the change (absolute or percent change) from Baseline to Week 24, treatment comparisons will be performed using a mixed model repeated measures approach with treatment, randomization stratification factors, visit, and treatment by visit interactions included as fixed effects. The Baseline value will be included as a covariate, and an unstructured variance-covariance matrix will be assumed. Calculation of the dependent variable (change from Baseline) for each patient at each visit will be calculated based on the visit windows specified in Section 4.3.5. Based on this model, the least squares mean at Week 24 will be compared between treatment groups and summarized along with the corresponding 95% CIs for treatment difference. In addition, summary statistics (mean change or mean % change) will be graphically presented as appropriate.

Binary Endpoints

For endpoints evaluating proportions, treatment comparisons will be performed using a stratified Cochran-Mantel-Haenszel test as appropriate with the randomization stratification factors as strata. Point estimates and 2-sided 95% CIs for treatment differences in proportions will be provided.

Descriptive statistics (point estimates and corresponding 95% CIs) will be provided by treatment group and visit as appropriate for all secondary endpoints.

Responder rate by visit (at Week 4, Week 8, Week 12, Week 16, and Week 20) will be derived in a similar fashion to the derivation of responder rate at Week 24/EOT. The missing data handling rules described in Section 7.3.5 for deriving responder status at Week 24/EOT will be applied similarly at Weeks 4, 8, 12, 16, and 20.

7.4.3. Derivation of Amenorrhea-Related Endpoints

Determination of Amenorrhea

Rules for determining amenorrhea in the treatment period is defined as those who meet 1 of the following requirements for 2 consecutive visits (approximately 56 consecutive days). Patients will be deemed to have amenorrhea during a visit window according to the following rules:

- No feminine product returned due to reported amenorrhea in 2 consecutive visits

OR

- No feminine product returned due to other reasons or feminine product collection with a negligible observed MBL volume coupled with other data indicating infrequent non-cyclic bleeding/spotting as described in Table 10.

Missing responses for menstrual bleeding questions in the eDiary will be treated as “No Bleeding” if eDiary compliance rate is > 70%.

Table 10: Rules for Determining Amenorrhea by Visit

Feminine Product Collection (KCAS) [a]	Supporting Data	
	Menstruation Status eCRF	**eDiary**
No feminine product collection due to reported amenorrhea	No menses start/stop dates reported	N/A
No feminine product collection due to other reasons	Per instructions for non-cyclic bleeding patterns, menses start date is reported but no menses stop date reported	• Data indicating infrequent, non-cyclic bleeding/spotting defined as bleeding/spotting with feminine product use for no more than 3 consecutive days and no more than 5 days bleeding total per visit window • eDiary entry rate > 70%
Feminine product collection with negligible observed MBL volume defined as <5 mL	Full or partial menses start and stop dates	• Data indicating infrequent, non-cyclic bleeding/spotting defined as bleeding/spotting with feminine product use for no more than 3 consecutive days and no more than 5 days bleeding total per visit window • eDiary entry rate > 70%

Abbreviations: eCRF, electronic case report form; eDiary, electronic diary; MBL, menstrual blood loss; N/A = not applicable.

[a] There is no requirement for feminine product return rate, as the determination of amenorrhea is based on the eDiary response.

Amenorrhea During the Last 35 Days of Treatment

Patients with amenorrhea over the last 35 days of treatment are defined as those who meet the definition of amenorrhea. A patient's amenorrhea status will also be summarized at Weeks 8, 12, 16, and 20. If a patient does not return for her Week 24/EOT visit, the eDiary responses for the last 35 days of treatment will be evaluated. If the criteria for infrequent, non-cyclic bleeding or spotting as indicated in Table 10 is met and the criteria for amenorrhea is met at the prior visit, the patient will be categorized as amenorrheic at Week 24/EOT. At all other timepoints, patients who do not return for a specific visit will be assigned as not amenorrheic at that visit.

Time to Amenorrhea

Time to amenorrhea is defined as the weeks from date of first dose of study drug to the start date of the amenorrhea window. Time to sustained amenorrhea will also be estimated and plotted using the Kaplan-Meier method.

The start date of amenorrhea is defined as the last feminine product collection date prior to start of amenorrhea. For example, if a patient's feminine product was collected at her Week 4 visit

and MBL volume for this cycle did not indicate amenorrhea, and the patient reported amenorrhea on Week 8 and 12 visits, then time to start amenorrhea will be defined as starting on the date of feminine product collection for Week 4. Patients who are determined to have amenorrhea at Week 4 and Week 8 will use their Week 4 feminine product collection date as start date of amenorrhea. Patients without an event will be censored at the last assessment date prior to the last dose of the study drug.

Sustained Amenorrhea Rate by Visit

A patient's sustained amenorrhea status will be summarized at Weeks 8, 12, 16, 20, and 24, based on her time to achieving and maintaining amenorrhea until the date of last study drug dose as shown in Table 11. For example, at Week 8, a patient is considered to have achieved sustained amenorrhea status if her amenorrhea started before Week 8 and was observed every visit thereafter until the last dose of the study treatment. The proportion of patients with sustained amenorrhea will be summarized by visit. If a patient met the criteria for sustained amenorrhea but discontinues from the study, this subject's amenorrhea status will be carried forward to the Week 24 visit.

Table 11: Sustained Amenorrhea Rate by Visit

	Amenorrhea Window	
Time Point	Start	End
Week 8	Determined amenorrhea at Week 4	Amenorrhea observed at Week 8 and was observed at every visit thereafter until and including the last dose of study treatment
Week 12	Determined amenorrhea at Week 8	Amenorrhea observed at Week 12 and was observed at every visit thereafter until and including the last dose of study treatment
Week 16	Determined amenorrhea at Week 12	Amenorrhea observed at Week 16 and was observed at every visit thereafter until and including the last dose of study treatment
Week 20	Determined amenorrhea at Week 16	Amenorrhea observed at Week 20 and was observed at every visit thereafter until and including the last dose of study treatment
Week 24	Determined amenorrhea at Week 20	Amenorrhea observed at Week 24

7.4.4. Derivation of Patient-Reported Outcome

7.4.4.1. Numerical Rating Scale Score for Pain Associated with Uterine Fibroids

Patients completed daily eDiaries including assessment of uterine fibroid-associated pain by the Numerical Rating Scale (NRS). Patients rated their worst pain in the last 24 hours caused by their uterine fibroids on a scale from 0 to 10, with 0 indicating no pain and 10 indicating pain as bad as you can imagine. The maximum NRS score for pain at Week 24/EOT is calculated as the maximum NRS score during the last 35 days on study treatment. If any NRS scores for pain during the last 35 days on study treatment are missing, the maximum score will be calculated as the maximum of all non-missing scores. Baseline NRS score for uterine fibroid-associated pain is defined as the maximum NRS score from the 35 days of data collected prior to randomization.

Proportion of women who achieve a *maximum* NRS score for pain associated with uterine fibroids over the last 35 days of treatment that is at least a 30% reduction from Baseline will be summarized in the subset of women with a maximum pain score ≥ 4 during the 35 days prior to randomization (subset). In addition, for the subset, mean maximum NRS scores will be provided by treatment and visit. Maximum NRS score for each patient at a visit is defined as the highest NRS score reported in the visit window specified in Table 2.

7.4.4.2. UFS-QoL Score

Calculation of UFS-QoL Symptom Severity Scale Score

To calculate the Symptom Severity Scale score, a summed score is created for the items listed below and then the formula below the table is used to transform raw scores to a normalized score with a range of possible values from 0 to 100. This provides Symptom Severity Scale scores, where higher scores are indicative of greater symptom severity and lower scores indicate lower symptom severity.

Scale	Sum Item Values	Lowest and Highest Possible Raw Scores	Possible Raw Score Range
Symptom Severity	Sum 1 – 8	8, 40	32

Formula for Transformation of Symptom Severity Raw Scores ONLY:

Transformed Score = [(Actual raw score – lowest possible raw score)/(Possible raw score range)] * 100

Calculation of UFS-QoL Bleeding and Pelvic Discomfort Scale Score

The UFS-QoL Bleeding and Pelvic Discomfort (BPD) Scale has been derived from the UFS-QoL Symptoms Scale; the derivation and validation of this new scale can be found in Appendix 3. The new scale consists of the following three symptoms proximal to uterine fibroids:

- Heavy bleeding during your menstrual period (Q1)
- Passing blood clots during your menstrual period (Q2)
- Feeling tightness or pressure in your pelvic area (Q5)

To calculate the score for the BPD Scale, a summed score of the items listed below is created and then the formula below the table is used to transform the raw score to a normalized score. This provides BPD Scale scores, where higher score values are indicative of greater symptom severity and lower scores will indicate minimal symptom severity (high scores = bad).

Sub-Scale	Sum Item Values	Lowest and Highest Possible Raw Scores	Possible Raw Score Range
Q1, Q2 and Q5	Sum 1,2,5	3, 15	12

Formula for Transformation of BPD Raw Scores ONLY:

Transformed Score = [(Actual raw score – lowest possible raw score)/(Possible raw score range)] * 100

On the basis of transformed score for BPD Scale, change from Baseline in the transformed score for BPD Scale at Week 24 will be defined as an alpha-protected key secondary endpoint comparing the relugolix + E2/NETA group with the placebo group. The proportion of patients who are responders (defined as meeting a meaningful change threshold from Baseline in the BPD Scale) at Week 24 on the transformed score for the BPD Scale will be compared between the two treatment arms (the relugolix + E2/NETA group with the placebo group) using a stratified Cochran-Mantel-Haenszel test, as appropriate. The proposed responder threshold is a 20-point change. Details in the determination of the meaningful change in the BPD Scale can be found in Appendix 4.

As a descriptive assessment on robustness of the responder analysis, a plot of the cumulative distribution function (CDF) will be provided for each treatment group to display the change from Baseline to Week 24 in the transformed score for BPD Scale on the x-axis and cumulative percentage of patients experiencing up to that change on the y-axis.

Calculation of Other UFS-QoL Scale Scores and UFS-QoL Total Score

For the other UFS-QoL scales (concern, activities, revised activities, energy/mood, control, self-conscious, and sexual function), a summed score of the items listed below is created for each individual scale. To calculate the UFS-QoL total score, the values for each individual scale are summed. Using the formula below the table, all raw scores are transformed to normalized scores. Higher scores are indicative of better health-related quality of life (high = good).

For endpoints evaluating a single question, the raw score is used in the analysis. The activity and revised activity domain scores will be summarized by treatment group.

Scale	Sum Item Values	Lowest and Highest Possible Raw Scores	Possible Raw Score Range
Concern	9+15+22+28+32	5, 25	20
Activities	10+11+13+19+20+27+29	7, 35	28
Revised activities	11+13+19+20+27	5,25	20
Energy/mood	12+17+23+24+25+31+35	7, 35	28
Control	14+16+26+30+34	5, 25	20
Self-conscious	18+21+33	3, 15	12
Sexual function	36+37	2, 10	8
HRQL TOTAL	Sum of 6 Subscale Scores[a]	29, 145	116

Abbreviations: HRQL, health-related quality of life.

[a] HRQL Total includes following scales: concern, activities, energy/mood, control, self-conscious, and sexual function.

Formula for Transformation of Raw Scores of Other Scale Scores ONLY:

Transformed Score = [(Highest possible score – Actual raw score)/(Possible raw score range)] * 100

For revised activities, the proportion of patients who are responders (defined as meeting a meaningful change from Baseline in the revised activity score) at Week 24 will be analyzed similarly to that for the change in BPD Scale score between the two treatment arms (relugolix + E2/NETA and placebo). The proposed responder threshold is a 20-point increase. Details of the determination of the meaningful change in the Revised Activities Scale score can be found in Appendix 5.

Missing Items

For any scale analyses, if < 50% of the scale items are missing, the scale should be retained using the mean scale score of the items present. If ≥ 50% of the items are missing, no scale score should be calculated; the subscale score will be considered missing.

7.4.4.3. Patient Global Assessment

The PGA for function and symptoms will be evaluated using a 5-point response scale (eg, absent, mild, moderate, severe, and very severe). To calculate change from Baseline to Week 24, the following numerical scores will be assigned to each response level:

Response Scale (Function)	Response Scale (Symptoms)	Numerical Score
No limitation at all	Not severe	1
Mild limitation	Mildly severe	2
Moderate limitation	Moderately severe	3
Quite a bit of limitation	Very severe	4
Extreme limitation	Extremely severe	5

For each item, the count and proportion of improvement by level or at least one level will be tabulated by treatment group and by visit. The denominator for the proportion will be based on the number of patients who provided non-missing responses to the items.

7.4.4.4. Menorrhagia Impact Questionnaire

The Menorrhagia Impact Questionnaire items 3 and 4 will be evaluated using the 5-point response scales (Not at all, Slightly, Moderately, Quite a bit, and Extremely) to assess level of improvement from Baseline to Week 24.

For each item, the count and proportion of improvement by level will be tabulated by treatment group and by visit. The denominator for the proportion will be based on the number of patients who provided non-missing responses to the items.

7.5. Exploratory Efficacy Endpoints

The following exploratory endpoints will be assessed for both comparisons the relugolix + E2/NETA group with the placebo group and the relugolix + delayed E2/NETA group with the placebo group:

- Change from Baseline to Week 24 in the EQ-5D-5L Scale score
- Change from Baseline to Week 24 in EQ-5D-5L visual analogue score.

7.5.1. Exploratory Efficacy Analyses

Analysis methods previously described for primary and secondary efficacy endpoint analyses will be used for the analysis of these endpoints.

8. PHARMACOKINETIC AND PHARMACODYNAMIC ANALYSES

Plasma relugolix, plasma NET, and serum E2 trough concentrations will be listed and summarized by study, treatment group (Group A, B, or C), and visit.

Serum pharmacodynamic data (LH, FSH, E2, and progesterone) will be listed and summarized using descriptive statistics (including raw and change from Baseline) by study, treatment group (Group A, B, or C), and visit.

For pharmacodynamic assessment, the number and percentage of patients with individual E2 concentration values < 10 pg/mL, 10 to < 20 pg/mL, 20 to < 50 pg/mL, and ≥ 50 pg/mL and individual progesterone concentration values < 1 ng/mL, 1 to 5 ng/mL, and ≥ 5 ng/mL will be summarized by treatment group (Group A, B, or C) and visit.

Scatter plots with LOESS smoothing lines for MVT-601-3001 and MVT-601-3002 separately will be used to examine the relationship between mean plasma relugolix trough concentration at the given time point (collected between 18 and 30 hours after the previous dose) and the following pharmacodynamic concentrations:

- Week 12 serum LH, FSH, E2, and progesterone (separately for Groups A and B);
- Week 24 serum LH, FSH, E2, and progesterone (separately for Groups A and B, and Groups A and B combined).

In addition, the PK data from this study will be combined with PK data from other studies to define a population PK model, which will be reported separately. Exposure-response analyses of the primary efficacy endpoint and safety will be conducted to assess the effect of relugolix exposure on outcomes. The analysis plan for population PK and exposure-response analyses will be specified in a separate document.

9. SAFETY ANALYSES

Unless otherwise specified, safety analyses will be conducted using the safety population according to the treatment received by the patients.

9.1. Adverse Events

Adverse events will be collected from the time of the first dose of study drug through the safety follow up visit approximately 30 days after the last dose of study drug (the end of treatment period), or the date of initiation of another investigational agent or hormonal therapy or surgical intervention or entering extension study, whichever occurs first. Serious adverse events reported to the investigator after the safety reporting period should be reported to the sponsor if the investigator assesses the event as related to study drug.

The severity of all treatment-emergent adverse events will be evaluated by the investigator based on the National Cancer Institute's Common Terminology Criteria for Adverse Events (CTCAE) and will be coded to preferred term and system organ class using MedDRA 22.0 or higher.

A treatment-emergent adverse event is defined as any adverse event that occurs after administration of the first dose of study drug.

Adverse event summaries will be based on treatment-emergent adverse events, unless otherwise specified. Adverse events occurring prior to administration of any study drug will be listed and flagged in by-patient listings.

The following tabular summaries that include the number and percentage of patients will be provided:

- Overview of adverse events;
- All adverse events;
  - By SOC and PT;
  - By decreasing frequency of PT;
  - By SOC, PT, and maximum severity;
  - Study drug-related per investigator by SOC and PT;
  - By time to onset, SOC and PT;
- Grade 3 or above adverse events;
  - By SOC and PT;
  - By decreasing frequency of PT;
  - Study drug-related per investigator by SOC and PT;
- Grade 2 or above adverse events;
  - By SOC and PT;
  - By decreasing frequency of PT;
  - By SOC, PT, and maximum severity;

  - Study drug-related per investigator by SOC and PT;
- Adverse events leading to study drug withdrawal;
  - By SOC and PT;
  - By decreasing frequency of PT;
- Adverse events leading to dose interruption;
  - By SOC and PT;
  - By decreasing frequency of PT;
- Adverse events resulting in fatal outcome;
  - By decreasing frequency of PT;
- Serious adverse events;
  - By SOC and PT;
  - By decreasing frequency of PT;
  - By SOC, PT, and maximum severity;
  - By SOC, PT, and relationship to study drug;
- Adverse events of clinical interest (ALT or AST ≥ 3 x ULN);
  - By SOC, PT, and maximum severity;
  - By decreasing frequency of PT.

Additionally, adverse event categories defined in Table 12 will be summarized by decreasing frequency of PT.

9.1.1. Relationship to Study Drug

Adverse events will be classified as "related" to study treatment if the relationship was rated by the investigator as possibly related or probably related. Adverse events related to any study drug (relugolix or placebo and E2/NETA or placebo) will be considered as related to study drug.

9.1.2. Severity of Adverse Event

Grade 2 or above adverse events will be summarized by SOC, PT, and/or maximum severity, relationship to study treatment.

9.1.3. Serious Adverse Event

Serious adverse events will be summarized by SOC, PT, and/or maximum severity, relationship to study treatment.

The data handling conventions for and the definition of a serious adverse event are discussed in this section. All deaths during the study, including the post treatment follow-up period, and deaths that resulted from a process that began during the study, should be included in the

analysis. For more details, deaths occurring during the following time periods or under the following conditions should be considered:

- Deaths occurring during participation in any study, or during any other period of drug exposure
- Deaths occurring after a patient leaves a study, or otherwise discontinues study drug, whether or not the patient completes the study to the nominal endpoint, if the death:
  - Is the result of a process initiated during the study or other drug exposure, regardless of when it actually occurs; or
  - Occurs within a time period that might reflect drug toxicity for a patient leaving a study or otherwise discontinuing drug. For drugs with prompt action and relatively short elimination half-lives, 4 weeks is a reasonable time period. For drugs with particularly long elimination half-lives or drug classes with recognized potential to cause late occurring effects, deaths occurring at longer times after drug discontinuation should be evaluated.

9.1.4. Adverse Event Leading to Withdrawal of Study Drug

Adverse events leading to withdrawal of study drug are those adverse events collected from the adverse event CRF pages with “drug withdrawn” as the action taken with study drug.

Adverse events with “drug withdrawn” as action taken due to any one of the components of study drug will be considered as leading to withdrawal of study drug.

9.1.5. Adverse Events Leading to Dose Interruption

Adverse events leading to dose interruption are those adverse events collected from the adverse event CRF pages with “drug interrupted” as their action taken with study drug.

Adverse events with “drug interrupted” as action taken due to any one of the components of study drug will be considered as leading to dose interruption.

9.1.6. Adverse Events Resulting to Fatal Outcome

Adverse events resulting in a fatal outcome are those adverse events collected from the adverse event pages with “fatal” as their outcome.

The fatal events, if any, will be provided in a by-subject listing.

9.1.7. Adverse Event Categories

In addition, adverse event categories defined in Table 12 will be summarized by decreasing frequency of PT under each safety population. Incidence of vasomotor symptoms by 12 weeks will be compared between relugolix Group A and relugolix Group B. Comparative statistics (such as p-values, 95% CIs, risk ratio) will be provided. Vasomotor symptoms throughout the studies will be summarized by SOC, PT, and maximum severity.

Table 12: Constitution of Adverse Event Categories

Category	Search Criteria
Bone health events	Osteoporosis/Osteopenia SMQ (broad) Fracture (custom SMQ): All preferred terms including the term "fracture," excluding "Tooth fracture" and "Fracture of penis"
Hepatic disorders	Drug-related hepatic disorders – comprehensive SMQ (narrow)
Metabolic disorders	Dyslipidemia SMQ (broad) Hyperglycemia/new onset diabetes mellitus SMQ (narrow)
Vasomotor symptoms	The following 5 preferred terms will be included: Hyperhidrosis; Feeling hot; Hot flush; Night sweats; Flushing
Mood disorders	MedDRA Depression and Suicide/Self-Injury SMQ (broad)

Abbreviations: HLT, High-Level Term; MedDRA, Medical Dictionary for Regulatory Activities; SMQ, Standardised MedDRA Query.

9.2. Laboratory Data

Laboratory parameters, including chemistry and hematology panels, specified as per protocol, and collected from the central laboratory will be tabulated and presented in by-patient listings. Urinalysis and hepatitis virus serological test results will be provided in by-patient listing only.

The National Cancer Institute CTCAE Grading Scale with numeric component will be used to categorize toxicity grade for laboratory parameters (CTCAE v5.0, dated 17 Nov 2017). Parameters that have criteria available for both low and high values (eg, hypercalcemia for a high value of calcium and hypocalcemia for a low value of calcium) will be summarized for both criteria (low and high). Patients will only be counted once for each criterion. The same patient can be counted for both criteria if she has laboratory values meeting each criterion. Shift tables will be provided for each gradable parameter to summarize Baseline toxicity grade versus worst post-Baseline toxicity grade. For laboratory parameters that are not gradable by the CTCAE, a shift table based upon the normal range (low, normal, and high) will be provided for each parameter to summarize the Baseline versus worst post-Baseline results.

Boxplots of laboratory values over time will be plotted for key laboratory parameters. These laboratory parameters include, but are not limited to, hematology (hemoglobin, platelets, leukocytes, neutrophils), creatinine, glomerular filtration rate, and hepatic function panel (alanine

aminotransferase [ALT], aspartate aminotransferase [AST], alkaline phosphatase [ALP], and total bilirubin).

The change from Baseline to each post-Baseline study visit will be presented by treatment group for each laboratory test in both tables and figures.

The number and proportion of patients with liver test elevations will be presented by treatment group. Liver test elevations are assessed by using post-Baseline results for ALT, AST, ALP, and total bilirubin based on the definitions presented in Table 13.

Table 13: Categories of Liver Test Elevations

Laboratory Test	Category
ALT or AST	ALT or AST > ULN - < 3xULN ALT or AST ≥ 3x to < 5x ULN ALT or AST ≥ 5x to < 8x ULN ALT or AST ≥ 8x to < 10x ULN ALT or AST ≥ 10 to < 20x ULN ALT or AST ≥ 20x ULN
Total bilirubin	Total bilirubin > 2 × ULN
ALT or AST and total bilirubin	ALT or AST ≥ 3 × ULN + total bilirubin > 2 × ULN
ALT or AST, total bilirubin, and ALP	ALT or AST ≥ 3 x ULN + total bilirubin > 2 × ULN + ALP < 2 × ULN

Abbreviations: ALP, alkaline phosphatase; ALT, alanine aminotransferase; AST, aspartate aminotransferase; ULN, upper limit of normal.

The number and percentage of patients with concurrent (defined as measurements on the same day) ALT or AST ≥ 3 times ULN and total bilirubin > 2 times ULN will also be presented.

9.3. Other Safety Analyses

9.3.1. Electrocardiograms

ECG interval results and changes from Baseline will be summarized descriptively for each scheduled visit in both tables and figures using data provided by and read by central reading.

A categorical analysis of corrected QT interval using Fridericia's calculation (QTcF) intervals will also be performed for each scheduled visit and for the maximum post-Baseline value. The number and percentage of patients in each QTcF interval category (< 450 msec, 450 to 480 msec, 481 to 500 msec, and > 500 msec) will be summarized. Categories of changes from Baseline (≥ 30 msec and ≥ 60 msec) will be summarized as well.

ECG intervals will be presented in by-patient listing. Overall ECG assessments performed by local reading will also be listed.

9.3.2. Visual Acuity

Visual Acuity Score at Baseline and at each scheduled post-Baseline assessment time point will be presented in a by-patient listing.

9.3.3. Vital Signs

Blood pressure (systolic and diastolic), heart rate, and BMI will be summarized at Baseline and each subsequent scheduled assessment by treatment group. Change from Baseline will be calculated and presented for each parameter at all scheduled post-Baseline assessment time points in both tables and figures. All vital sign data will also be provided in by-patient listings.

Potentially clinically significant abnormalities in vital signs are defined in Table 14, and they will be summarized by using post-Baseline values that meet the defined criteria. Potentially clinically significant abnormalities will also be flagged in by-patient listings.

Table 14: Categories of Potentially Clinically Significant Abnormalities in Vital Signs

Parameter	Category
Systolic blood pressure	≥ 140 mmHg ≥ 180 mmHg
	≤ 90 mmHg
	Increase of ≥ 20 mmHg from Baseline Decrease of ≥ 20 mmHg from Baseline
Diastolic blood pressure	≥ 90 mmHg ≥ 105 mmHg
	≤ 50 mmHg
	Increase of ≥ 15 mmHg from Baseline Decrease of ≥ 15 mmHg from Baseline
Heart rate	≥ 120 bpm
	< 45 bpm
	Increase of ≥ 15 bpm from Baseline Decrease of ≥ 15 bpm from Baseline

Abbreviations: bpm, beats per minute; mmHg, millimeters of mercury.

9.3.4. Endometrial Biopsy

Primary diagnosis of endometrial biopsy assessment will be summarized at Baseline and at scheduled assessment by treatment group. All endometrial biopsy data will also be provided in a by-patient listing.

Primary diagnosis from pathologist evaluation will be categorized by medical monitor's review in Table 15 and will be summarized using frequencies and percentages, summarized for each treatment group. All endometrial biopsy data will also be provided in by-patient listings.

Table 15: Categories of Primary Diagnosis in Endometrial Biopsies

Normal-Proliferative	• Weakly proliferative • Proliferative • Disordered proliferative
Normal-Secretory/Menstrual/Mixed	• Secretory • Menstrual • Progestational/Decidulized/Mixed
Normal-Atrophic or Minimally Stimulated	• Atrophic • Indeterminate/Inactive
Hyperplasia	• Simple hyperplasia without atypia • Simple hyperplasia with atypia • Complex hyperplasia without atypia • Complex hyperplasia with atypia
Carcinoma	—
Inadequate	—
Missing	—
Additional Diagnosis (Other reported finding)	• Reactive/Inflammatory • Polyp • Metaplasia • Glandular and/or Stromal Breakdown

9.3.5. Bone Mineral Density

Corrected BMD data will be used for analysis as determined by the central radiology laboratory in the 3 prespecified anatomical locations: lumbar spine (L1–L4), total hip, and femoral neck.

BMD at Baseline, Week 12 and Week 24 visits will be summarized descriptively by treatment group and each anatomical location. Percentage changes from Baseline along with 95% CIs of mean percentage changes will be also summarized by treatment group and anatomical location. Mean percentage change from Baseline with its corresponding 95% CI will be plotted by visit, treatment group, and anatomical location.

To support the inclusion of E2/NETA in the treatment regimen, the safety endpoint of mean percent change from Baseline in BMD at the lumbar spine at Week 12 will be analyzed using pooled data from the two replicate studies (MVT-601-3001 and MVT-601-3002) with a formal

comparison of the relugolix + E2/NETA group (Group A) versus the relugolix + delayed E2/NETA group (Group B) (details in the Integrated Summary of Safety Statistical Analysis Plan).

In addition, the difference of percentage change from Baseline between treatment groups (relugolix + E2/NETA group versus the relugolix + delayed E2/NETA group at 12 weeks, relugolix + E2/NETA versus placebo group at 12 and 24 weeks, and relugolix + delayed E2/NETA group versus placebo group at 12 weeks) will be summarized at each visit by anatomical location along with the corresponding 95% CIs.

To account for participants whose BMD assessment may have been obtained outside of the protocol-specified window (Week 12 ± 3 weeks, Week 24 ± 3 and 4 weeks), a sensitivity analysis by visit will be conducted that includes all women who underwent DXA at both time points, regardless of whether the image was procured during the prespecified time window.

A mixed-effects model with repeated measures will be used to describe treatment effect on BMD at 12 and 24 weeks. The model will have treatment group, age at Baseline, visit, Baseline BMD value, stratification factors (geographic region and menstrual blood loss volume), race (African American versus Other), and BMI at Baseline as fixed effects using an unstructured variance-covariance matrix. Least square means on each anatomical location will be presented and plotted at each visit with associated 95% CIs. Categorical representation of percentage change from Baseline to 12 and 24 weeks of treatment will be presented by the number and proportion of patients who had BMD declines of ≤ 2%, >2% to 3%, > 3% to 5%, > 5% to 8%, and > 8% by treatment group and anatomical location. The 95% CIs will be provided for the respective proportions.

Categorical changes from Baseline in overall BMD (defined as lumbar spine and total hip) also will be assessed at 12 and 24 weeks. Femoral neck evaluates a smaller area of bone mass than the total hip and is prone to lower precision in the measurement (ISCD Official Positions, 2015; Leslie, 2007). Since femoral neck BMD may be associated with discordant readings compared with the total hip or lumbar spine due to technical considerations, it will not add meaningful interpretation of overall BMD changes in response to treatment.

Z-scores will be summarized by treatment group, visit, and anatomical location with descriptive statistics including 95% CIs, and the number and percentage of patients with a Z-score < -2.0 will be presented by treatment group, visit, and anatomical location.

BMD percentage changes from Baseline will also be summarized by intrinsic factors (eg, age, race, body mass index) and extrinsic factors (eg, geographic region).

9.3.6. Bleeding Pattern

Bleeding patterns will be summarized at Week 24/EOT by treatment group. Three bleeding patterns will be considered: amenorrhea (see Section 7.4.3), cyclic bleeding, and irregular bleeding. Patients with the cyclic bleeding pattern are those who do not meet the definition of amenorrhea and do meet the following conditions:

- 3 to ≤ 12 days of menstruation duration per eDiary at Week 24/EOT window (see Section 7.3.3)

- No more than 2 days of gap of no bleeding (per eDiary) within the menstruation duration.

Patients with the irregular bleeding pattern are those who do not meet the definitions of cyclic bleeding or amenorrhea. The number (and percent) of patients and mean number of bleeding days will be provided by treatment group for each bleeding pattern.

For patients with cyclic or irregular bleeding pattern, the number (and percent) of patients with observed MBL volume falling into the following bleeding intensity groups will be provided:

- **Spotting/negligible bleeding:** MBL volume < 5 mL
- **Light:** MBL volume 10 - 50 mL
- **Moderate:** MBL volume >50 to ≤80 mL
- **Heavy:** MBL volume > 80 mL

For each bleeding intensity category, the mean number of bleeding days will be summarized.

10. REFERENCES

2015 International Society for Clinical Densitometry (ISCD) Official Positions – Adult (https://www.iscd.org/official-positions/2015-iscd-official-positions-adult/; accessed 30 Apr 2019).

Bachowski G, Borge D, Brunker P, Eder A, Fialkow L, Friday J, et al. A Compendium of Transfusion Practice Guidelines. American Red Cross; 2017 p. 81. Report No.: 3rd Edition.

Leslie WD, Lix LM, Tsang JF, Caetano PA. Single-site vs Multisite Bone Density Measurement for Fracture Prediction. Arch Intern Med 2007; 167 (15): 1641-7

Man L, Tahhan HR. Body surface area: a predictor of response to red blood cell transfusion. JBM. 2016 Sep;Volume 7:199–204.

Ratitch B, Lipkovich I, O'Kelly M. Combining analysis results from multiply imputed categorical data. PharmaSUG 2013. Paper SP03.

Rubin D. The calculation of posterior distributions by data augmentation: comment: a noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: the SIR algorithm. J Am Stat Assoc. 1987 Jun; 82(398), 543-546.

Stewart EA, Owens C, Duan WR, Gao J, Chwalisz K, Simon JA. Elagolix alone and with add-back decreases heavy menstrual bleeding in women with uterine fibroids. Poster presented to American College of Obstetricians and Gynecologists Annual Clinical and Scientific Meeting 2017, San Diego, CA, May 6-9, 2017.

Sullivan TR, White IR, Salter AB, Ryan P, Lee KJ. Should multiple imputation be the method of choice for handling missing data in randomized trials? Stat Methods Med Res. 2018 Sep; 27(9):2610-2626.

von Hippel PT. How many imputations do you need? A two-stage calculation using a quadratic rule. Sociol Methods Res. 2018 Jan.

APPENDICES

APPENDIX 1. SUMMARY OF SECONDARY ENDPOINT ANALYSES

Secondary Endpoints	Analysis Population	Statistical Method/Test	Declare Statistical Significance [a]	Time Points of Summary	Summary Statistics
Key Secondary Efficacy Endpoints with Alpha Protection					
Proportion of women who achieve amenorrhea over the last 35 days of treatment	mITT	Cochran-Mantel-Haenszel test	P < 0.05	Week 24	Frequency and percentages
% change from Baseline to Week 24 in MBL volume	mITT	Mixed-effects model	P < 0.05	Week 24	LS means for % change
Proportion of women with a hemoglobin ≤10.5 g/dL at Baseline who achieve an increase of > 2 g/dL from Baseline at Week 24	Subset of mITT	Cochran-Mantel-Haenszel test	P < 0.05	Week 24	Frequency and percentages
Change from Baseline to Week 24 in the UFS-QoL Bleeding and Pelvic Discomfort Scale score, a sub-scale of the UFS-QoL Symptom Severity Scale	mITT	Mixed-effects model	P < 0.05	Week 24	LS means for change
Proportion of patients with a maximum NRS score ≤ 1 during the 35 days before the last dose of study drug in the subset of women with a maximum NRS score ≥4 for pain associated with uterine fibroids during the 35 days prior to randomization	Subset of mITT	Cochran-Mantel-Haenszel test	P < 0.05	Week 24/EOT	Frequency and percentages
% change from Baseline to Week 24 in uterine fibroid volume	mITT	ANCOVA model	P < 0.05	Week 24	LS means for % change
% change from Baseline to Week 24 in uterine volume	mITT	ANCOVA model	P < 0.05	Week 24	LS means for % change
Other Secondary Efficacy Endpoints					
Time to achieve MBL volume of < 80 mL AND at least a 50% reduction from Baseline MBL volume as measured by the alkaline hematin method	mITT	Log-rank test/KM method	P < 0.05	Monthly	KM estimates at Week 12 and 24, KM plots, median time to response
Time to achieve amenorrhea	mITT	Log-rank test/KM method	P < 0.05	Monthly	KM plots, median time to response

Secondary Endpoints	Analysis Population	Statistical Method/Test	Declare Statistical Significance [a]	Time Points of Summary	Summary Statistics
Time to sustained amenorrhea	mITT	Log-rank test/KM method	P < 0.05	Monthly	KM plots, median time to response
Proportion of women in the relugolix Group A versus the placebo Group C who achieve an MBL volume of < 80 mL AND at least a 50% reduction from Baseline MBL volume at Week 4, Week 12, Week 16, and Week 20	mITT	No comparison		*at Week 4, Week 12, Week 16, and Week 20*	Descriptive
Sustained amenorrhea rate by visit	mITT	Cochran-Mantel-Haenszel test	P < 0.05	Monthly	Frequency and percentages
Proportion of women with a hemoglobin below the lower limit of normal at Baseline who achieve an increase of ≥ 1 g/dL from Baseline at Week 24	mITT	Cochran-Mantel-Haenszel test	P < 0.05	Monthly	Frequency and percentages
Change (absolute and %) from Baseline to Week 24 in hemoglobin for women with a hemoglobin ≤ 10.5g/dL at Baseline	mITT	Mixed-effects model	P < 0.05	Monthly	LS means for % change
Proportion of women who achieve a maximum Numerical Rating Scale score for uterine fibroid-associated pain over the last 35 days of treatment that is at least a 30% reduction from Baseline in the subset of women with a maximum pain score ≥ 4 during the 35 days prior to randomization	Subset of mITT	Cochran-Mantel-Haenszel test	P < 0.05,	Week 12, 24	Frequency and percentages
Mean maximum NRS scores over time	Subset of mITT	Descriptive		Monthly	Means
Proportion of responders who had meaningful reduction of >20 points from Baseline to Week 24 in UFS-QOL Bleeding and Pelvic Discomfort Scale (Q1, Q2 and Q5)	mITT	Cochran-Mantel-Haenszel test	P < 0.05,	Week 12, 24	Frequency and percentages
Proportion of responders who had meaningful increase of > 20 points from Baseline to Week 24 in UFS-QOL revised activities	mITT	Cochran-Mantel-Haenszel test	P < 0.05,	Week 12, 24	Frequency and percentages

Secondary Endpoints	Analysis Population	Statistical Method/Test	Declare Statistical Significance [a]	Time Points of Summary	Summary Statistics
Change from Baseline to Week 24 in impact of uterine fibroids based on the UFS-QOL revised activities domain	mITT	Mixed-effects model	P < 0.05	Week 12, 24	LS means for change
Change from Baseline to Week 24 in impact of uterine fibroids based on the UFS-QOL activities domain	mITT	Mixed-effects model	P < 0.05	Week 12, 24	LS means for change
Change from Baseline to Week 24 in the interference of uterine fibroids with physical activities based on UFS-QOL Q11	mITT	Mixed-effects model	P < 0.05	Week 12, 24	LS means for change
Change from Baseline to Week 24 in the interference of uterine fibroids with social activities based on UFS-QOL Q20	mITT	Mixed-effects model	P < 0.05	Week 12, 24	LS means for change
Change from Baseline to Week 24 in embarrassment caused by uterine fibroids based on UFS-QOL Q29	mITT	Mixed-effects model	P < 0.05	Week 12, 24	LS means for change
Change from Baseline to Week 24 in the UFS-QoL Symptom Severity Scale score	mITT	Mixed-effects model	P < 0.05	Week 12, 24	LS means for change
Change from Baseline to Week 24 in the UFS-HRQL total score	mITT	Mixed-effects model	P < 0.05	Week 12, 24	LS means for change
Change in PGA for uterine fibroid related function from Baseline to Week 24	mITT	Mixed-effects model	P <0.05	Monthly	LS means for absolute and change
Change in PGA for uterine fibroid symptoms from Baseline to Week 24	mITT	Mixed-effects model	P < 0.05	Monthly	LS means for absolute and change
Proportion of patients achieving improvement in PGA for uterine fibroid symptoms from Baseline to Week 24	mITT	Cochran-Mantel-Haenszel test	P < 0.05	Monthly	Frequency and percentages
Proportion of patients achieving improvement in PGA for uterine fibroid related function from Baseline to Week 24	mITT	Cochran-Mantel-Haenszel test	P < 0.05	Monthly	Frequency and percentages
Safety Related Endpoints					
% Change from Baseline to Week 12 in BMD (pooled data)	Safety population	Mixed-effects model Relugolix Group A vs B	P < 0.05,	Week 12	LS means Diff (95%CI)

Secondary Endpoints	Analysis Population	Statistical Method/Test	Declare Statistical Significance [a]	Time Points of Summary	Summary Statistics
% Change from Baseline in BMD	Safety population	Mixed-effects model Relugolix Group A vs Placebo at 12/24 weeks; Relugolix Group B vs Placebo at 12 weeks		Week 12, 24	LS means Diff (95%CI)
Exploratory Secondary Efficacy Endpoints					
Change from Baseline to Week 24 in the Menorrhagia Impact Questionnaire Score for physical activities	mITT	Descriptive		Monthly	Frequency and percentages
Change from Baseline to Week 24 in the Menorrhagia Impact Questionnaire Score for social and leisure activities	mITT	Descriptive		Monthly	Frequency and percentages

Abbreviations: KM, Kaplan-Meier; LS, least squares; mITT, modified intent-to-treat; NRS, Numerical Rating Scale; Q, question; UFS-HRQL, Uterine Fibroid Scale – Health-related Quality of Life.

[a] P-values are two-sided.

APPENDIX 2. DERIVATION AND PSYCHOMETRIC EVALUATION OF A UFS-QOL BLEEDING AND PELVIC DISCOMFORT SCALE

The BPD Scale was derived from the Symptom Severity Scale of the Uterine Fibroid Symptom and Health-Related Quality of Life (UFS-QoL). The BPD Scale consists of three items proximal to uterine fibroids that are experienced by most patients, (ie, heavy bleeding during the menstrual period [Question 1], passing blood clots during the menstrual period [Question 2], and feeling tightness or pressure in the pelvic area [Question 5]).

The aim of this appendix is to describe the derivation and psychometric testing process of the BPD Scale. Results of the analyses in this appendix are summarized in Appendix 3 and will be included in the Patient-Reported Outcomes dossier to be submitted at the time of filing for the uterine fibroids registration program.

Exploratory factor analysis and subsequent confirmatory factor analysis were conducted to assess and confirm the factor structure of the Symptom Severity Scale of the UFS-QoL, using data from a phase 2 study of relugolix in uterine fibroids (TAK-385/CCT-001), as well as pooled, blinded data from one-third of patients in the phase 3 studies (MVT-601-3001 and MVT-601-3002). Respective analyses are described in Section 2.1. Based on the results, the factor(s) reflecting symptoms proximal to uterine fibroids and experienced by most patients with uterine fibroids were selected for further psychometric testing.

The psychometric properties of the new scale were assessed using the same pooled, blinded data from the two phase 3 studies of relugolix in uterine fibroids (MVT-601-3001 and MVT-601-3002). These analyses are described in Section 2.2. The blinded data consists of the first third of patients (approximately n = 260) enrolled into the two pivotal studies who have completed the patient global assessment (PGA) for symptoms and the UFS-QoL at Baseline and at Week 24. Of note, for the analyses specified in Section 2.2, only data at Baseline and Week 12 were used; the Week 24 data was used in the responder analyses described in Appendix 3.

2.1. Development of the Bleeding and Pelvic Discomfort Scale Using Phase 2 and Phase 3 Data

From a review of the eight items in the Symptom Severity Scale of the UFS-QoL, it was apparent that the scale consists of different constructs/dimensions. Therefore, the factor structure of the Symptom Severity Scale was assessed, initially using data from the phase 2 study TAK-385/CCT-001 (n = 216).

Of note, in the TAK-385/CCT-001 phase 2 study, the UFS-QoL with a one-month recall period was applied, whereas the UFS-QoL with a three-month recall period is used in the phase 3 studies (MVT-601-3001 and MVT-601-3002). Therefore, confirmatory factor analysis and final psychometric testing of the chosen factor was conducted using blinded phase 3 data (see Section 2.2).

Exploratory Factor Analysis

The exploratory factor analysis was done on phase 2 data to identify the underlying constructs by the most parsimonious factor structure of the eight items in the Symptom Severity Scale. Identification of the number of factors was based on the following criteria:

- Items with primary factor loading > 0.4;
- Factors with large eigenvalues considered as common factors using Kaiser criterion (Kaiser, 1960).

A scree plot was used as a supplemental tool to decide on the number of factors in the final model.

Confirmatory Factor Analysis

Once the number of factors was identified, a confirmatory factor analysis was conducted using blinded, pooled phase 3 data to confirm the factor structure. Only patients who completed the Baseline and Week 24 PGA for symptoms and UFS-QoL assessments were included in this analysis. Model fit was assessed based on the following:

- The goodness of fit as measured by χ^2 and Goodness of Fit Index; a Goodness of Fit Index > 0.9 is considered acceptable;
- The Comparative Fit Index was used to determine the acceptability of the model fit of the discrepancy function adjusted sample size; a Comparative Fit Index > 0.9 (Hu, 1995) was considered an acceptable fit;
- The root mean square error of approximation was used to determine the acceptability of model fit of the square root of the discrepancy between the sample covariance matrix and the model covariance matrix; the root mean square error of approximation had to be < 0.06 (Browne, 1993) to be considered an acceptable fit;
- P-value > 0.05.

Once the final factor structure was identified, the factor reflecting items proximal to uterine fibroids and experienced by almost all patients with uterine fibroids were selected for further evaluation. Of note, this was the BPD Scale.

2.2. Psychometric Analyses Based on Phase 3 Data

The same pooled, blinded data from the first third of patients enrolled in either of the two phase 3 studies (MVT-601-3001 or MVT-601-3002) was used for the psychometric analyses of the BPD Scale. The objective was to psychometrically evaluate the new scale in terms of item performance, reliability, validity, and ability to detect change. Of note, for the analyses specified in this section, only data at Baseline and Week 12 were used. The following analyses were performed:

Item Level Analysis Assessing Ceiling and Floor Effects:

- A descriptive summary of the eight items in the UFS-QoL Symptom Severity Scale at Baseline was provided to examine item distributions and ceiling/floor effects. Low ceiling effects (< 20%) and higher floor effects (> 20%) were expected at Baseline

due to symptom severity of patients with uterine fibroids enrolled in the phase 3 studies.

Internal Consistency:

Internal consistency reliability was assessed for the BPD Scale at Baseline and Week 12 by calculating Cronbach's alpha. Generally, a Cronbach's alpha coefficient (α) ≥ 0.7 indicates an acceptable level of internal consistency.

Item Performance:

- Intercorrelation of items that contribute to the BPD Scale by means of item-total correlation was determined.
- Item discrimination index was assessed for each item based on 1) the BPD Scale scores at single time points, and 2) the change from Baseline to Week 12 in the BPD Scale score to determine the degree to which individual items were able to discriminate between less and more severe patients (Cappelleri, 2014).

Known-Groups Validity:

- Known-groups validity was assessed based on groups defined by Baseline PGA for symptoms severity (five levels). Descriptive statistics of the BPD Scale will be provided for each severity level.

Ability to Detect Change:

Evidence that the new scale can identify differences in scores over time in individuals or groups who have changed with respect to the measurement concept will be investigated by providing the following descriptive statistics:

- Within person change from Baseline to Week 12 in each item on the BPD Scale
- Standardized effect size statistic (SES) for change from Baseline to Week 12 in each item scale. The ability to detect change will be judged based on Cohen's recommendations: small change (SES = 0.20), moderate change (SES = 0.50), and large change (SES = 0.80).

2.3. References

Browne MW, Cudeck R. Alternative ways of assessing model fit. In: Bollen KA, Long JS (eds), Testing structural equation models (Vol. 154, pp. 136-162). 1993. Newbury Park, CA: Sage Focus Editions.

Cappelleri JC, Bushmakin AG. Interpretation of patient-reported outcomes. Stat Methods Med Res. 2014;23(5), 460–483.

Hu LT, Bentler PM. Evaluating model fit. In: Hoyle RH (ed), Structural equation modeling: concepts, issues, and applications (pp. 76-99). 1995. Thousand Oaks, CA, US: Sage Publications, Inc.

Kaiser HF. The application of electronic computers to factor analysis. Educ Psychol Meas. 1960;20:141-151.

APPENDIX 3. DERIVATION AND VALIDATION OF THE UFS-QOL BLEEDING AND PELVIC DISCOMFORT SCALE

Results described in this appendix are based on the analyses described in Appendix 2.

3.1. Development of the Bleeding and Pelvic Discomfort Scale Using Exploratory and Confirmatory Factor Analysis

Exploratory factor analysis was conducted on data from the phase 2 study TAK-385/CCT-001 study (n = 216) and the arising factor structure was assessed in a confirmatory factor analysis using data from the phase 3 studies MVT-601-3001 and MVT-601-3002.

3.1.1. Exploratory Factor Analysis Using Phase 2 Data

Exploratory factor analysis results revealed a two-factor solution based on the Kaiser criterion (eigenvalues > 1) and factor loading > 0.40 criteria specified in the analysis plan (see Appendix 2). Factor 1 and Factor 2 had eigenvalues of 3.394 and 1.196, respectively (Table 3.1-1). Three items were found to load adequately onto Factor 1 with loadings greater than 0.40: Item 1 (Heavy Bleeding during Your Period), Item 2 (Passing Blood Clots during Your Period), and Item 5 (Feeling Tightness or Pressure in Pelvis; see Table 3.1-2). Two items loaded onto Factor 2 with loadings larger than the prespecified level: Item 6 (Frequent Urination in Daytime) and Item 7 (Frequent Nighttime Urination). Item 8 (Feeling Fatigued) showed a loading value on Factor 1 just below the prespecified threshold (0.399) and showed evidence of cross-loading with the Factor 2 (0.288). An additional factor with a moderate eigenvalue (0.62) was considered based the scree plot (Figure 3.1-1) and factor loadings of its associated items (Item 3: Fluctuation in Duration of Menstruation, 0.416; Item 4: Fluctuation in Length of Monthly Cycle, 0.995; Table 3.1-2).

Overall the results show support for a seven-item three-factor model. Due to multi-factor loading, Item 8 (Feeling Fatigued) remains a single-item symptom and is not scored as part of any factor.

Figure 3.1-1: Scree Plot and Variance Explained for UFS-QoL Symptom Severity Scale Factors in TAK-385/CCT-001

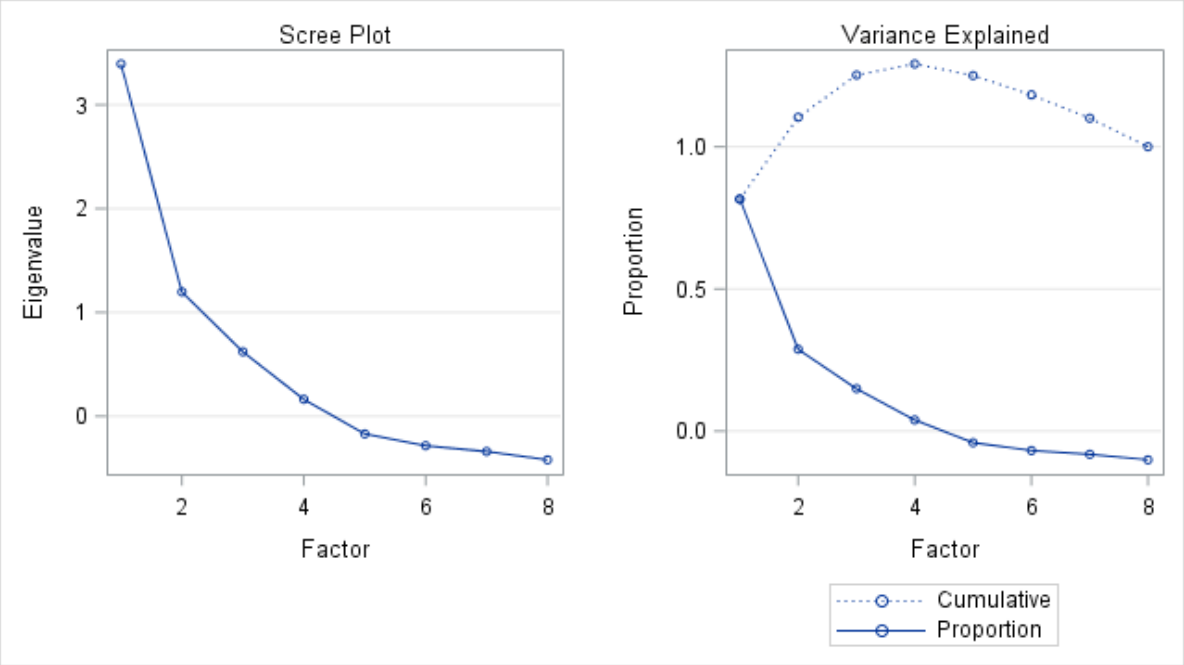


Table 3.1-1: Exploratory Factor Analysis for the UFS-QoL Symptom Severity Scale in TAK-385/CCT-001

Item	Eigenvalue	Difference	Proportion	Cumulative
1	3.394	2.198	0.816	0.816
2	1.196	0.576	0.288	1.104
3	0.620	0.458	0.149	1.253
4	0.162	0.332	0.039	1.292
5	-0.170	0.114	-0.041	1.251
6	-0.284	0.057	-0.068	1.183
7	-0.341	0.079	-0.082	1.101
8	-0.419	—	-0.101	1.000

Table 3.1-2: Factor Loadings for the UFS-QoL Symptom Severity Scale in TAK-385/CCT-001

Items		Factor1	Factor2	Factor3
Q2	Passing blood clots during your period	**0.763**	0.105	0.073
Q1	Heavy bleeding during your period	**0.759**	0.091	0.123
Q5	Feeling tightness or pressure in pelvis	**0.467**	0.175	0.167
Q8	Feeling fatigued	0.399	0.288	0.078
Q6	Frequent urination in daytime	0.114	**0.965**	0.069
Q7	Frequent nighttime urination	0.212	**0.630**	0.013
Q4	Fluctuation in length of monthly cycle	0.039	0.092	**0.995**
Q3	Fluctuation in duration of menstruation	0.178	0.003	**0.416**

Extraction method: maximum likelihood. Rotation method: orthogonal.

3.2. Development of the Bleeding and Pelvic Discomfort Scale Using Confirmatory Factor Analysis Based on Phase 3 Data

The exploratory factor structure arising from the phase 2 data was assessed using data from the phase 3 studies MVT-601-3001 and MVT-601-3002.

Analyses were based on pooled, blinded data from the first one third of patients enrolled in the two phase 3 studies of relugolix in uterine fibroids (MVT-601-3001 and MVT-601-3002), who completed the patient global assessment of symptoms (PGA) and the Uterine Fibroid Symptom and Health-Related Quality of Life (UFS-QoL) at Baseline and at Week 24.

3.2.1. Confirmatory Factor Analysis using Phase 3 Data

A confirmatory factor analysis was completed using blinded data from one third of phase 3 patients. The acceptance criteria of the confirmatory factor analysis were prespecified as a Goodness of Fit Index > 0.90 and a Comparative Fit Index > 0.90, a root mean square error of approximation < 0.06 and a non-significant p-value to show that the null-hypothesis that the data fits the three-factor model was not rejected (Table 3.2-1).

Factor loadings for the seven-item three-factor model supported the three-factor solution proposed by the exploratory factor analysis in the above described analyses using phase 2 data. Results indicated that the three-factor model, excluding item 8, had a Goodness of Fit Index and a Comparative Fit Index of 1.00 and a root mean square error of approximation of 0.00 (90% CI = 0.00-0.02). The test of model fit returned a p-value of 0.9394. The null hypothesis that the data fit the model was not rejected (see Table 3.2-1). Under this model, Item 5 (Feeling Tightness or Pressure in Pelvis) also cross-loaded onto Factor 2, assessing urinary symptoms.

Table 3.2-1: Confirmatory Factor Analysis of the UFS-QoL Symptom Severity Scale without Item 8: Model Fit Statistics at Baseline (MVT-601-3001 and -3002)

Model Fit Statistics[a]					
Model		**CFI**	**RMSEA (90%CI)**	**GFI**	**P-value**
3-Factor Model (7-item)		1.000	0.000 (0.00-0.02)	1.000	0.9394
Factor Loading[b]					
			Factor1	**Factor2**	**Factor3**
Q1	Heavy bleeding during your period		**0.7314**	0.2672	0.2024
Q2	Passing blood clots during your period		**0.7620**	0.1503	0.2099
Q3	Fluctuation in duration of menstruation		0.3263	0.1861	**0.6909**
Q4	Fluctuation in length of monthly cycle		0.1689	0.1561	**1.0323**
Q5	Feeling tightness or pressure in pelvis		**0.4644**	**0.4657**	0.1965
Q6	Frequent urination in daytime		0.2503	**0.7727**	0.1300
Q7	Frequent night time urination		0.1553	**0.8605**	0.1538

Abbreviations: CFI, comparative fit index; CI, confidence interval; GFI, goodness of fit index; RMSEA, root mean square error approximation.

[a] Model fit statistics allow for assessment of the model appropriateness.

[b] Rotation Method: Orthogonal.

In order to further assess the performance of the Fatigue item, which was excluded following the exploratory factor analysis due to cross-loading, the confirmatory factor analysis was reconducted with the inclusion of this item in Factor 1. Results showed that the eight-item three-factor model had a Goodness of Fit Index of 0.996, a Comparative Fit Index of 1.00 and a root mean square error of approximation of 0.00 (90% CI = 0.00-0.05). The test of model fit returned a p-value of 0.8056. However, the results for Item 8 showed a cross-loading of this item at 0.417 on Factor 1 and 0.437 on Factor 2 (Table 3.2-2). This continued cross-loading supports the exclusion of this item in the scoring of any factor (Table 3.2-2).

Table 3.2-2: Confirmatory Factor Analysis of the UFS-QoL Symptom Severity Scale with Item 8 included: Model Fit Statistics at Baseline (MVT-601-3001 and 3002)

Model Fit Statistics[a]					
Model		**CFI**	**RMSEA (90%CI)**	**GFI**	**P-value**
3-Factor Model (8-item)		1.000	0.000 (0.00-0.05)	0.996	0.8056
Factor Loading[b]					
			Factor1	Factor2	Factor3
Q1	Heavy bleeding during your period		**0.732**	0.265	0.211
Q2	Passing blood clots during your period		**0.750**	0.150	0.226
Q3	Fluctuation in duration of menstruation		0.296	0.175	**0.767**
Q4	Fluctuation in length of monthly cycle		0.180	0.167	**0.932**
Q5	Feeling tightness or pressure in pelvis		**0.473**	**0.465**	0.206
Q6	Frequent urination in daytime		0.251	**0.757**	0.137
Q7	Frequent night time urination		0.150	**0.876**	0.156
Q8	Feeling fatigued		**0.417**	**0.437**	0.136

Abbreviations: CFI, comparative fit index; CI, confidence interval; GFI, goodness of fit index; Q, question; RMSEA, root mean square error of approximation.

[a] Model fit statistics allow for assessment of the model appropriateness.

[b] Rotation Method: Orthogonal.

3.3. Classical Test Theory Psychometric Analyses of the Bleeding and Pelvic Discomfort Scale Based on Phase 3 Data

Each of the above-described factor analyses showed that a seven-item three-factor solution was appropriate for the UFS-QoL Symptom Severity Scale. Following this confirmation, blinded psychometric appraisal of the measure was implemented to further understand the performance of the items and subscales of the UFS-QoL Symptom Severity Scale. For the item level analysis, all items were assessed. For subscale level analysis, the analysis was focused, primarily, on the evaluation of the Factor 1 – the Bleeding and Pelvic Discomfort (BPD) Scale. The BPD Scale was selected as the primary focus for further psychometric evaluation, as it presents clinical and patient-reported symptoms proximal to the disease and is associated with high symptom burden experienced by most patients.

Analyses were based on pooled, blinded data from the first one third of patients enrolled in the two phase 3 studies of relugolix in UF (MVT-601-3001 and MVT-601-3002) who completed the PGA for symptoms and the UFS-QoL at Baseline and at Week 24. Of note, for the analyses specified in this section, only data at Baseline and Week 12 were used.

3.3.1. Item Level Analysis of the UFS-QoL Symptom Severity Scale

UFS-QoL Symptom Severity Scale item responses were assessed for floor (highest possible severity) and ceiling effects (lowest possible severity). Overall, the measure showed no ceiling effects (response option 1, Table 3.3-1, demonstrating that the items have scope to capture

patient improvement in disease burden. A greater proportion of patients responded at floor level (response option 5; range =11.15 to 36.15%), which is expected at the start of a clinical trial. All response options for all items were used, showing a good coverage of the range of disease burden. When considering BPD Scale items, all items showed a range of responses that covered the response scale, with over 50% of patients reporting being a (very) great deal distressed by heavy bleeding during menstrual period (Item 1), passing blot clots during menstrual period (Item 2), and feeling of tightness or pressure in the pelvic area (Item 5).

Table 3.3-1: Summary of UFS-QoL Symptom Severity Scale Response at Baseline by Items in MVT-601-3001 and 3002

	Q1 (N = 260)		Q2 (N = 260)		Q3 (N = 260)		Q4 (N = 260)		Q5 (N = 260)		Q6 (N = 260)		Q7 (N = 260)		Q8 (N = 260)	
Response	n	(%)	n	(%)	n	(%)	n	(%)	n	(%)	n	(%)	n	(%)	n	(%)
1	4	(1.54%)	4	(1.54%)	44	(16.92%)	63	(24.23%)	21	(8.08%)	48	(18.46%)	54	(20.77%)	13	(5.00%)
2	15	(5.77%)	30	(11.54%)	48	(18.46%)	37	(14.23%)	24	(9.23%)	35	(13.46%)	53	(20.38%)	21	(8.08%)
3	53	(20.38%)	61	(23.46%)	66	(25.38%)	69	(26.54%)	57	(21.92%)	77	(29.62%)	64	(24.62%)	59	(22.69%)
4	101	(38.85%)	71	(27.31%)	64	(24.62%)	62	(23.85%)	96	(36.92%)	62	(23.85%)	55	(21.15%)	82	(31.54%)
5	87	(33.46%)	94	(36.15%)	38	(14.62%)	29	(11.15%)	62	(23.85%)	38	(14.62%)	34	(13.08%)	85	(32.69%)

Abbreviations: N, number of patients; n, number of patients in subset; Q, question.

3.3.2. Scale Level Analysis of the BPD Scale

3.3.2.1. Internal Consistency

Internal consistency was assessed for the BPD Scale at Baseline and Week 12. Reliability was acceptable at Baseline (> 0.70) and good at Week 12 (> 0.80; Table 3.3-2).

Table 3.3-2: Cronbach's Alpha Coefficient of BPD Scale by VISIT (MVT-601-3001 and 3002)

		Q1	Q2	Q3	
	n	Mean (SD)	Mean (SD)	Mean (SD)	**Alpha[a]**
Baseline	260	3.97 (0.95)	3.85 (1.09)	3.59 (1.18)	0.768
Week 12	258	2.75 (1.47)	2.69 (1.46)	2.64 (1.36)	0.882

Abbreviations: n, number of patients; Q, question; SD, standard deviation.
[a] Cronbach Coefficient Alpha

3.3.2.2. Item-to-Total Correlations

Item-to-total correlations were assessed to ensure that each item was associated with the BPD Scale score. Correlations demonstrate that each of the items have a strong relationship with the total score at Baseline and at Week 12 ($r > 0.50$) (Table 3.3-3). Correlations improved at Week 12, which represents a greater spread of the data across each item's five-point response scale, further supporting the relationship of these items to the BPD total score.

Table 3.3-3: Intercorrelation of Items in BPD Scale by Visit (MVT-601-3001 and 3002)

Question	**Baseline** **N = 260**	**Week 12** N = 258
Q1	0.670	0.802
Q2	0.620	0.845
Q5	0.533	0.674

Note: Intercorrelation calculated using Pearson's correlations.

3.3.2.3. Item Discrimination Indices

An item discrimination index was employed to assess the ability of each item to discriminate between high and low severity patients. At Baseline, the discrimination index represents each item's ability to differentiate patients on the BPD Scale scores at a single time point, and at Week 12, the discrimination index represents the ability to differentiate patients based on their level of change from Baseline to Week 12 in the BPD Scale score.

Results show that all items had a discrimination index above 0.60, demonstrating that BPD Scale items are able to discriminate between high- and low-severity patients both when assessing single time point scores and change over time (Table 3.3-4).

Table 3.3-4: Item Discrimination Index of BPD Scale (MVT-601-3001 and 3002)

	Q1	Q2	Q5
Baseline (n = 260)	0.815	0.954	0.923
Week 12 (n = 258)	0.915	0.986	0.836

Abbreviations: n, number of patients; Q, question.
Note: BPD scale upper/lower ranges**:** Upper = at least 65-point reduction, Lower = at most 10-point reduction.

3.3.2.4. Known-Groups Validity

A known-groups analysis assessed the descriptive BPD score and score ranges for patients stratified by level of severity reported on the PGA (symptoms). Results from the known-groups validity assessment show that mean and median BPD Scale scores increase monotonically in line with PGA symptom severity (Table 3.3-5).

3.3.2.5. Ability to Detect Change

The BPD Scale's ability to detect change was assessed though the difference in BPD Scale scores over time in patients who have changed with respect to the measurement concept as measured by the PGA (symptoms). For each PGA stratified group, within person change from Baseline to Week 12 and standardized effect size statistics (SES) for change over the same period were assessed. SES statistics judged were based on Cohen's recommendations (small change, 0.20; moderate change, 0.50; large change, 0.80).

Results showed that the mean change for improving PGA categories had a monotonically increasing pattern from patients who had a PGA change of 0 to patients who had a PGA improvement of -4 (Table 3.3-6). Worsening groups (PGA change of +1 or +2) had very low levels of mean change, with wide standard deviations around the mean due to the low sample size in these categories.

In line with expectations, the SES statistics for the improvement categories (PGA score change of -1 to -4) were large (> 0.80) compared to the moderate SES found in the patients who reported no change (PGA score change of 0; SES = 0.55).

Table 3.3-5: Summary Statistics of BPD Scale Score at Baseline by PGA (symptoms) Response (MVT-601-3001 and 3002)

	Baseline BPD Scale Score[a]						
Baseline PGA	**N**	**Mean**	**SD**	**Median**	**Q1, Q3**	**Min**	**Max**
1	7	53.57	28.81	58.33	25.00, 75.00	16.67	91.67
2	21	59.92	26.56	58.33	41.67, 75.00	8.33	100.00
3	96	62.33	21.18	66.67	41.67, 75.00	8.33	100.00
4	89	75.09	19.48	75.00	66.67, 91.67	16.67	100.00
5	47	83.51	16.53	91.67	75.00, 100.00	41.67	100.00

Abbreviations: BPD, bleeding and pelvic discomfort; max, maximum; min, minimum; N, number of patients; PGA, Patient Global Assessment; Q1, first quartile; Q3, third quartile; SD, standard deviation.
a Transformed Score.

Table 3.3-6: Summary Statistics of Change from Baseline BPD Scale Score to Week 12 by PGA (symptoms) Change from Baseline (MVT-601-3001 and 3002)

PGA Change Category[a]	N	Mean	SD	95% CI	Median	Q1, Q3	Min	Max	Effect Size[b]
-4	23	-48.19	(42.27)	(-66.47, -29.91)	-66.67	-83.33, 0.00	-100.00	25.00	-2.93
-3	50	-49.33	(33.16)	(-58.76, -39.91)	-54.17	-75.00, -25.00	-100.00	33.33	-2.41
-2	74	-27.70	(30.75)	(-34.83, -20.58)	-25.00	-41.67, 0.00	-91.67	25.00	-1.25
-1	48	-23.09	(28.57)	(-31.39, -14.79)	-16.67	-33.33, -8.33	-100.00	33.33	-1.01
0	39	-10.68	(20.32)	(-17.27, -4.10)	-8.33	-25.00, 0.00	-66.67	33.33	-0.55
1	14	1.79	(19.11)	(-9.25, 12.82)	-4.17	-16.67, 8.33	-16.67	33.33	0.07
2	6	-1.39	(29.54)	(-32.39, 29.61)	-12.50	-25.00, 16.67	-25.00	50.00	-0.05

Abbreviations: BPD, blood and pelvic discomfort; CI, confidence interval; max, maximum; min, minimum; N, number of patients; PGA, Patient Global Assessment; Q1, first quartile; Q3, third quartile; SD, standard deviation.
Note: Statistics calculated using transformed score of BPD scale.
[a] The PGA is a five-point, single item patient-reported outcomes tool that measures patient's symptoms. The PGA change category with -4 = Marked Improvement; 0 = No Change, +4 = Markedly Worse.
[b] Standardized effect sizes are calculated as the mean divided by the standard deviation.

3.4. Conclusions

The exploratory factor analysis offered support for a three-factor solution, which included factors assessing Bleeding and Pelvic Discomfort, Urinary Symptoms, and Fluctuation in Menstruation. The Fluctuations in Menstruation factor had an eigenvalue < 1 but had items that loaded at greater than 0.40 and made theoretical sense as a construct.

The exploratory factor analysis showed that Item 8, measuring fatigue, cross-loaded on two factors (Bleeding and Pelvic Discomfort and Urinary Symptoms). Since fatigue is a multidimensional concept that can assess impacts and/or symptoms concurrently, it was not included in the final factor structure. Confirmatory factor analysis on the seven-item three-factor solution provided support for the exploratory factor structure; however, Item 5 cross-loaded between the BPD and Urinary Symptoms factors in this analysis. As Item 5 (Feeling Tightness or Pressure in Pelvis) is a proximal symptom of uterine fibroids, this item was retained as part of the BPD factor.

To ensure that fatigue was not being inappropriately excluded from the three-factor structure, an additional confirmatory factor analysis was conducted with fatigue included within the BPD factor. The inclusion of fatigue in this model continued to show the expected cross-loading of this item. This analysis confirmed that the multidimensional concept of fatigue was not suitable for inclusion in the BPD factor.

The BPD factor, which assesses symptomology most proximal to the disease, was further assessed through classical test theory psychometric evaluation. The results showed that the items of the BPD Scale work cohesively to inform the total score of the measure, and adequately distinguish between severities. At a score level, descriptive statistics were able to support the construct validity and responsiveness of the BPD Scale through showing a monotonic improvement in BPD Scale score in line with patient self-reported improvement on the PGA (symptoms). Additionally, by showing that the items of the BPD Scale perform well together, the psychometric results help to further support the inclusion of the cross-loading Item 5 on the BPD Scale.

APPENDIX 4. APPROACH TO ESTIMATING THE RESPONDER THRESHOLD OF THE UFS-QOL BLEEDING AND PELVIC DISCOMFORT SCALE

The Uterine Fibroid Symptom and Health-Related Quality of Life Bleeding and Pelvic Discomfort (UFS-QoL BPD) Scale includes the following items:

During the previous 3 months, how distressed were you by:

- Heavy bleeding during your menstrual period;
- Passing blood clots during your menstrual period;
- Feeling tightness or pressure in your pelvic area.

Response options include:

- Not at all;
- A little bit;
- Somewhat;
- A great deal;
- A very great deal.

The summary score of the three items included in the UFS-QoL BPD Scale ranges from 0 to 100, where a higher score indicates a higher level of distress and a lower score indicates a lower level of distress.

Change from Baseline to Week 24 in the BPD Scale score is an alpha-protected key secondary endpoint of the pivotal studies (MVT-601-3001 and MVT-601-3002) to evaluate the treatment benefit of relugolix + E2/NETA (Group A) compared with placebo (Group C). Additionally, a responder analysis will be performed between the two groups with respect to proportion of patients who have achieved a meaningful reduction from Baseline to Week 24 in BPD Scale score. This appendix describes the approach used to derive the responder threshold, including both the quantitative and supportive qualitative methods and the respective results.

The meaningful change threshold is the smallest reduction in the BPD Scale score that is considered meaningful by patients (Cohen, 1988; Crosby, 2003; Revicki, 2008; Cappelleri, 2014; Coon, 2018). The magnitude of a meaningful change threshold depends on the magnitude of the correlation between the BPD Scale change score and the Patient Global Assessment (PGA) of symptom severity (anchor) change and the variability of change on the BPD Scale by improvement categories on the PGA of symptom severity (described in Section 4.2.2). Several anchor-based methods will be used; however, the primary analysis will be a measure of central tendency for each improvement category (see Section 4.2.3). Anchor-based methods will use data collected on:

- The BPD Scale score at Baseline and Week 24; and
- The PGA of symptom severity score at Baseline and Week 24.

Results from the anchor-based analyses will be supported by qualitative data collected in a patient interview study (MVT-601-037), a sub-study of the phase 3 trials, in which patients from

selected sites in the United States (US) provided feedback on what they considered to be a meaningful change on the BPD Scale and the PGA of symptom severity (described in Section 4.2.4).

4.2. Statistical Analyses Plan for Estimation of the Responder Threshold

4.2.1. Anchor and Its Correlation with UFS-QoL Endpoint

The PGA of symptom severity uses a five-point verbal rating scale and asks the patient:

> "How severe were your uterine fibroids symptoms, such as heavy bleeding over the last four weeks?"
>
> Response options include:
>
> - Not severe;
> - Mildly severe;
> - Moderately severe;
> - Very severe;
> - Extremely severe.

The categorical change from Baseline to Week 24 in PGA of symptom severity score will be derived, leading to nine possible outcomes ranging from +4 (denoting worsening) to -4 (denoting improvement). The change in PGA of symptom severity at Week 24 will be used as the anchor (see Table 4.2-1).

4.2.2. Target Anchor Category

The target anchor category is the anchor category that represents the minimum meaningful change and is used as the starting point to identify potential candidates for a meaningful change threshold. For the two pivotal studies, the target anchor category will be a one-point category improvement on the PGA of symptom severity score (see Table 4.2-1), as this is typically considered as a minimal clinical important difference on a five-point Likert scale.

Table 4.2-1: Change in PGA of Symptom Severity as Anchor

Anchor	Anchor Change Category	Potential Target Anchor Change Category (To Be Used for Estimation of Meaningful Change Threshold)
Change in PGA of symptom severity	-4, -3, -2, -1 (improvement), 0 (same), +1, +2, +3, +4 (worsening)	-1-category change (improvement)

Abbreviations: PGA = patient global assessment.

4.2.3. Anchor-Based Methods

To determine the meaningful change threshold for the reduction in USF-QoL BPD Scale score, the anchor-based analyses described below will be performed.

The category (or point) change in PGA of symptom severity score will be used as the anchor to classify patients into response groups depending on their level of symptom severity change from Baseline to Week 24 (see Table 4.2-1). Uncollapsed, categorical change on the PGA will range from +4 to -4. Collapsed, categorical change will be considered based on the distribution of change categories on the PGA of symptom severity. Usually the collapsing occurs on the tails with extreme worsening (+4) or improvement (-4).

Among the anchor-based analyses described below, the within-group analysis will be primary and other analyses (including between-group analysis) are supportive.

4.2.3.1. Correlation with Anchor

Correlation between the categorical change on the PGA of symptom severity score and the change in the BPD Scale score will be evaluated at Week 24, using blinded pooled data from the first third of the enrolled patients from the two pivotal studies who have completed Week 24 visits and have the corresponding PGA of symptom severity data available (denoted as the "threshold determination analysis set"). Polyserial correlation coefficient will be used with a criteria value of > 0.30 indicating meaningful correlation (Crosby, 2003; Revicki, 2008; Cappelleri, 2014; Coon, 2018).

4.2.3.2. Within-Group Meaningful Change

Magnitude of change from Baseline to Week 24 in BPD Scale score will be calculated within each anchor category group. Changes in BPD Scale scores are negative for symptom reductions and positive for symptom increases.

Descriptive statistics (*n*, mean change, median change, 25th and 75th percentiles, standard deviation [SD], confidence interval [CI], and standardized effect size [SES]) will be reported for the changes in BPD Scale scores by anchor category. The SES will be calculated for each level of anchor category group by dividing the mean change score of BPD Scale from Baseline by the Baseline SD of the anchor category group. The impact of treatment will be judged based on Cohen's recommendations (1988): small change (SES = 0.20), moderate change (SES = 0.50), and large change (SES = 0.80). Significance associated with within-patient change will be evaluated using paired t-tests on the change in BPD Scale score separately for each level of improvement on the anchor.

4.2.3.3. Supportive Analysis of Between Group Meaningful Change Using Analysis of Variance

Analysis of variance (ANOVA) will be used to determine whether a difference in mean change scores from Baseline to Week 24 on the UFS-QoL BPD Scale exists between the categorical change groups (or the collapsed groups, as appropriate). Providing there is a significant change in UFS-QoL BPD Scale scores between the (collapsed) anchor groups, the between-group differences will be explored. Any anchor group with at least 15 patients will be included in this analysis. An anchor group with < 15 patients (usually occurring on the tails with extreme

worsening [+4] or improvement [-4]) will be collapsed with its adjacent group as appropriate. Comparison of the anchor groups of interest between the target anchor (-1 change category) and the "0 change" category will be performed using a t-test. The statistically significant difference on the BPD Scale change scores corresponding to a 1-category change on the PGA of symptom severity can be used as supportive information for estimating the meaningful change threshold.

4.2.3.4. Visualizing Cumulative Distribution Function and Probability Distribution Function Plots by Anchor Category Group

Anchor-based meaningful change will also be evaluated using cumulative distribution function (CDF) plots utilizing the Kernel smoothing for all anchor category groups, based on cumulative change in UFS-QoL BPD Scale scores for all available changes from Baseline to Week 24. Specifically, the CDF plot for each anchor category displays the probability (presented on the y-axis) of patients who have achieved a given absolute change of X or less in BPD Scale score from Baseline to Week 24 for each point change along the range of possible absolute changes (from -100 [maximum reduction] to 0 [no change] to 100 [maximum increase]) expressed on the x-axis.

Similarly, the smooth probability density function (PDF) will also be plotted for each anchor category group over the range of absolute changes in BPD Scale scores. These probabilities are plotted on the y-axis, with the BPD Scale change score on the x-axis.

The CDF and PDF curves are delineated by anchor improvement category (from -4 to +4) displaying the center and separation between the curve for the target anchor group and the curve for the group reporting no change on PGA of symptom severity. It is expected that the CDF curves will not cross between the change category groups (eg, monotonic increase from no change to slightly improved and moderately improved).

4.2.4. Determining a Meaningful Change Threshold Using the Totality-of-Evidence Approach

The meaningful change threshold will be determined using the totality of evidence from the results of above quantitative anchor-based analyses; results from the interview study (MVT-601-037) will be used as supportive evidence.

The results of these analyses and proposed thresholds will be included into the Patient-Reported Outcome dossier to be submitted at the time of filing.

4.3. Results from Anchor-Based Analyses

4.3.1. Correlation of Change in BPD with PGA of Symptom Severity

Meaningful change for the UFS-QoL BPD Scale was derived based on anchor-based methods, supported by cumulative distribution function (CDF) and probability density function (PDF) curves. To assess the suitability of the selected anchor, PGA of symptom severity, a polyserial correlation was calculated between change on the PGA from Baseline to Week 24 and the change from Baseline to Week 24 on the BPD Scale. The change in the PGA was moderately correlated (r = 0.57) with the change on the BPD Scale (Table 4.3-1). Given that the PGA is less complex than the BPD scale, this result indicates that the PGA is a suitable anchor for the BPD Scale.

4.3.2. Improvement on BPD Scale by PGA Change Category

Uncollapsed changes on the PGA were used to determine minimal meaningful improvement on the BPD Scale (Table 4.3-1). Improvement on the BPD Scale increased monotonically for all the categories from "no change (0)" to "1-category improvement (-1)" to "2-category improvement (-2)" to "3 category improvement (-3)" with nonoverlapping 95% CIs for mean change of the groups. Table 4.3-1 shows further that a 1-category improvement (-1) is associated with a 27.31-point mean improvement in the BPD Scale score at Week 24 compared with Baseline, with a 95% CI [-35.42, -19.19], a large SES = -1.21, and a median improvement of 25.00 points.

Table 4.3-1: Summary of Change from Baseline to Week 24 in UFS-QoL BPD Scale by PGA for Symptom Severity Change Category (mITT Population)

PGA Change Category	N = 255	Change in BPD					Correlation between PGA Change and BPD Change[a]
		Mean (SD)	Median	95% CI	p-value[b]	SES[c]	
4-Category deterioration (+4)	0						0.57
3-Category deterioration (+3)	2	-12.50 (5.89)	-12.5	-65.44, 40.44	0.2048	-2.12	
2-Category deterioration (+2)	2	0.00 (11.79)	0	-105.89, 105.88	1.00	0.0	
1-Category deterioration (+1)	21	-10.32 (16.22)	-8.33	-17.70, -2.93	0.0086	-0.54	
0-Category deterioration (0)	47	-9.93 (23.09)	-8.33	**-16.71**, -3.15	0.005	-0.42	
1-Category improvement (-1)	47	**-27.31** (27.62)	-25.00	-35.42, **-19.19**	< 0.0001	-1.21	
2-Category improvement (-2)	68	-42.16 (25.71)	-41.67	-48.38, -35.93	< 0.0001	-1.93	
3-Category improvement (-3)	45	-61.85 (26.62)	-66.67	-69.85, -53.85	< 0.0001	-3.25	
4-Category improvement (-4)	23	-54.35 (32.65)	-66.67	-68.47, -40.23	< 0.0001	-4.12	

Abbreviations: BPD = bleeding and pelvic discomfort; CI = confidence interval; mITT = modified intent-to-treat; PGA = patient global assessment; SD = standard deviation; SES = standardized effect size.
mITT is used to calculate change from Baseline score at Week 24 and includes patients from the mITT population who have available change from Baseline data at Week 24.

[a] Polyserial correlation coefficient between change in BPD Scale and change in PGA of symptom severity.

[b] The p-value for each individual change group is derived from a paired (within-sample) t-test assessing the difference over time.

[c] SES is calculated as the mean divided by the SD of Baseline. SES is judged as small = 0.2, moderate = 0.5, and large = 0.8 (Cohen 1988).

Table 4.3-2 highlights that the difference between the "1-category improvement" and the "no change" groups (mean = -17.38 with a 95% CI of [-27.81, -6.94]) was statistically significant (p = 0.0013) and had a moderate SES = -0.736, which also supports the notion that patients interpreted these change categories as distinct.

Patients were able to distinguish between the PGA improvement categories, as demonstrated by the nonoverlapping CIs (in Table 4.3-2) for their UFS-QoL BPD Scale scores and as illustrated

by the clear separation between the CDF curves presented in Figure 4.3-1. Since statistically significant differences existed in patient responses on the BPD Scale between the "1-category improvement (-1)" option and the "no change" and "2-category improvement (-2)" options, a 1-category improvement on the PGA was considered a meaningful target anchor category for assessing the responder threshold on the BPD Scale. Although a 2-category improvement could have been considered for deriving the meaningful change threshold, such a threshold would not qualify as being the *minimum* threshold possible. Given the statistical difference between the 1- and 2-category improvements and the fact that patients were able to distinguish between the two response options (to be taken up shortly), the evidence supports using a 1-category improvement on the PGA for estimating the minimum meaningful change threshold. This decision is also supported by qualitative evidence generated from the Exit Interview study (see Section 4.2.4).

Table 4.3-2: Summary of Change from Baseline to Week 24 in BPD Scale Between Target Anchor (-1) and No Change (0) in PGA of Symptom Severity (mITT Population)

Anchor	Categorical Change	N	Mean Change from BL	SD	95% CI	p-value[a]	Baseline SD	SES
PGA	1-category improvement (-1)	47	-27.31	27.62	-35.42, -19.19		22.63	
	No change (0)	47	-9.93	23.09	-16.71, -3.15		23.61	
	Difference		-17.38	25.46	-27.81, -6.94	0.0013		-0.736[b] -0.790[c]

Abbreviations: ANOVA = analysis of variance; BL = Baseline; BPD = bleeding and pelvic discomfort; CI = confidence interval; mITT = modified intent-to-treat; PGA = patient global assessment; SD = standard deviation; SES = standardized effect size.

[a] The p-value is based on t-test for difference in mean change in BPD score between the 2 anchor groups (-1 and 0) from the ANOVA in which the +2, +3, and +4 groups were collapsed with the +1 group due to 0 or few patients in the respective groups.

[b] SES is calculated as the mean difference divided by the SD of Baseline for no change group. They are judged as small = 0.2, moderate = 0.5, and large = 0.8 (Cohen 1988).

[c] SES calculated as the mean difference divided by the standard deviation of Baseline for pooled from all categories (Glass 1976).

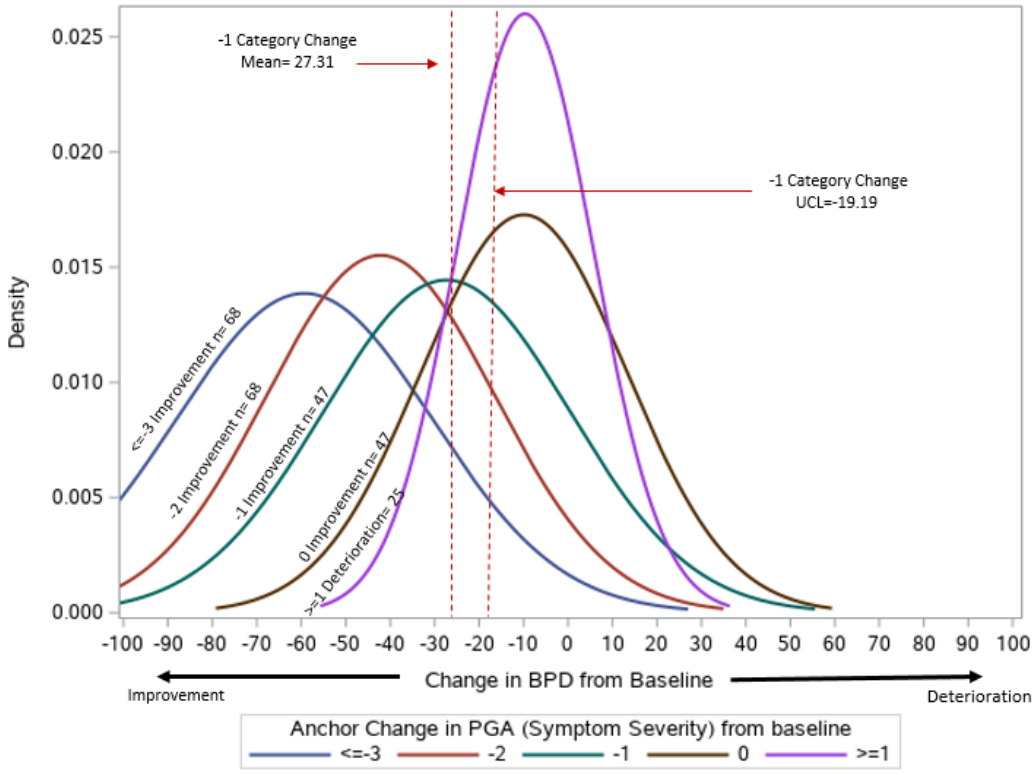
4.3.3. Estimation of Responder Threshold

Examination of the PDF curves, presented in Figure 4.3-1, indicates that the dispersion is roughly the same for the options between "> 3 category improvement" and "no change." The crossing of the "no change" and "1-category improvement" PDF curves at approximately -24 points (ie, a 24-point improvement on the BPD between Baseline and Week 24) indicates the meaningful change threshold is greater (less negative) than this value, because to the left of the value the "1-category improvement" was more probable than the "no change" curve. That is, to the left of this point (larger improvements) patients were more likely to be responders than to the right of this point. However, since the goal is to establish the minimum meaningful change threshold, the value -24 points is likely too conservative.

Using the mean or median values for measuring improvement in the BPD Scale would also yield estimates that are too conservative, because expected values do not necessarily constitute a *minimum* meaningful change threshold for patients. That is, nearly half the patients stratified in

the PGA "1-category improvement" who reported changes smaller than (to the right of) the mean or median on the BPD Scale would be classified as nonresponders by using the mean or median as the threshold despite of their reporting "1-category improvement." A less conservative, though still plausible estimate for the minimal meaningful change threshold is the upper bound of the 95% CI for mean change in the "1-category improvement" group. Its use will result in a smaller proportion of patients being classified as nonresponders in change on the BPD Scale than the expected value (ie, the mean). According to the uncollapsed anchor-based analysis (Table 4.3-1), this value is approximately -19 (ie, a 19-point improvement on the BPD Scale between Baseline and Week 24). Selection of this value is supported by the fact that the mean changes are statistically significantly different (Table 4.3-2) between "no change" and "1-category improvement" groups with clear separation of the respective 95% CIs for mean change. Of note, a value as low as -17 could also be selected, since it is less than the lower-bound 95% CI estimate of -16.71 for the "no change" group.

Figure 4.3-1: PDF of the Change in UFS-QoL BPD Scale by PGA Anchor Change Category (Collapsed)

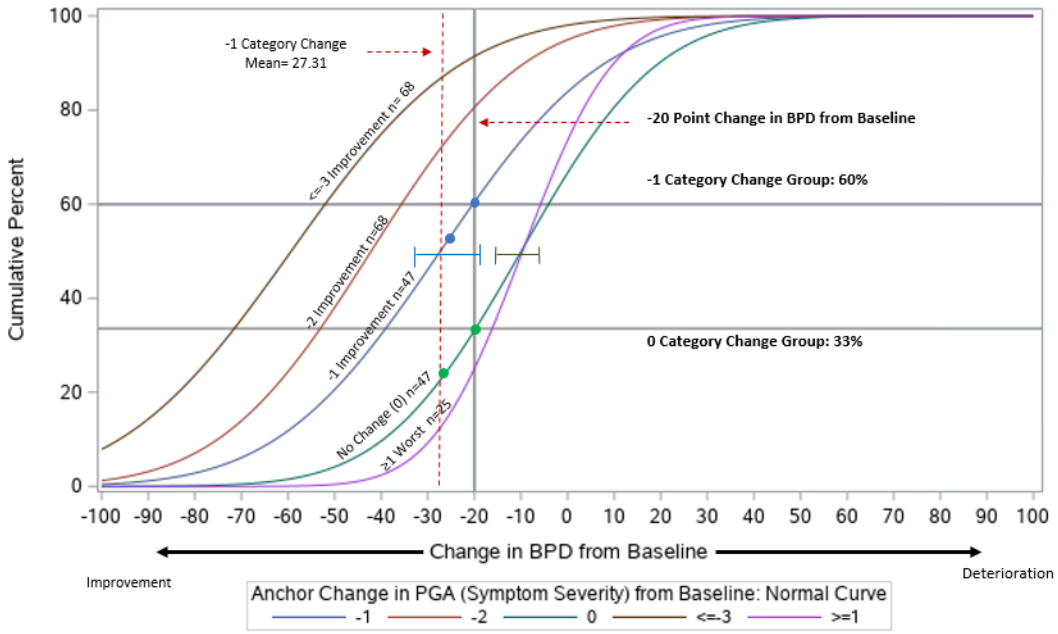


Abbreviations: BPD = bleeding and pelvic discomfort; PGA = patient global assessment; UCL = upper confidence limit.

Examination of the CDF curves for the potential minimum meaningful threshold value of -19 points on the BPD Scale allows one to estimate the cumulative percent of patients that would

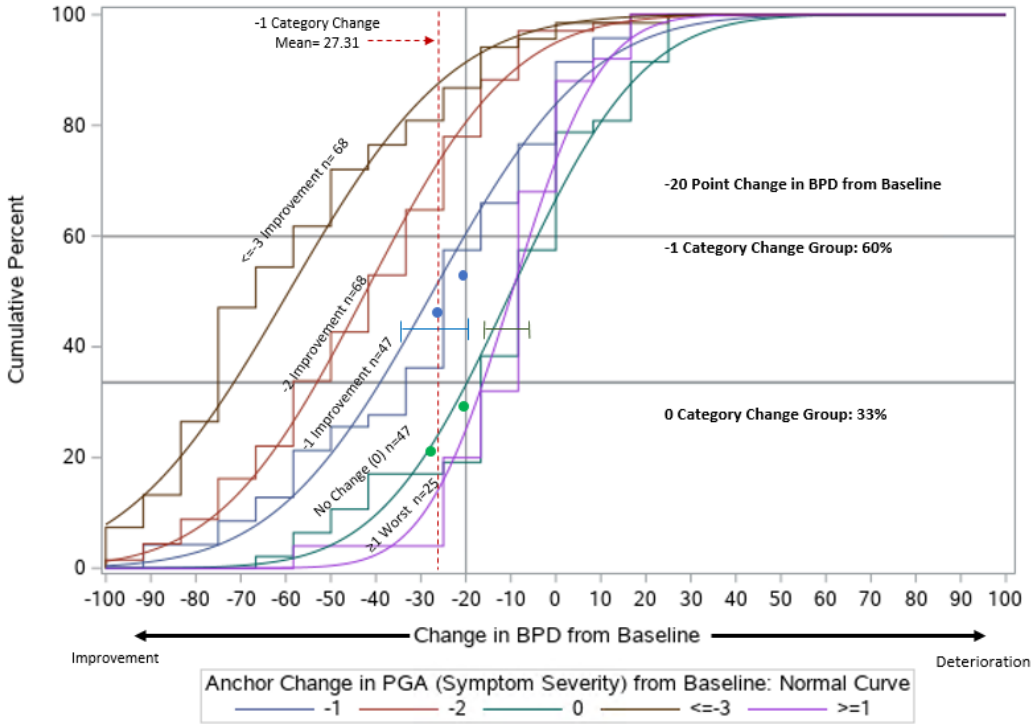
experience the improvement. As illustrated in Figure 4.3-2, approximately 35% of the "no change" group and 61% of the "1-category improvement" group experienced at least a 19-point improvement on the BPD Scale by Week 24. The high percent of patients in the "no change" group who improved on the BPD Scale by Week 24 indicates that setting the minimum meaningful change threshold at 19 points may be too liberal. The percent of misclassified responders can be improved by selecting a slightly larger value. Setting the minimum meaningful change threshold at 20-point improvement on the BPD Scale would decrease slightly the percent of misclassified responders for the "no change" group to 33% while decreasing slightly the percent of patients classified as responders to 60% for the "1-category improvement" group. As supportive information, the empirical CDFs were step-curves (reflecting the discrete nature of the BPD scores) are provided (Figure 4.3-3), indicating that smooth curves are reasonably close to the empirical CDFs.

Figure 4.3-2: Cumulative Distribution Function of Change at Week 24 in UFS-QoL BPD Scale Score by PGA Anchor Change Category (Collapsed)



Abbreviations: BPD = bleeding and pelvic discomfort; PGA = patient global assessment.

Figure 4.3-3: Empirical Cumulative Distribution Function of Change at Week 24 in UFS-QoL BPD Scale Score by PGA Anchor Change Category (Collapsed)



Abbreviations: BPD = bleeding and pelvic discomfort; PGA = patient global assessment.

4.4 Exit Interview Study Synthesis

4.4.1 Objectives

The objectives of the exit interviews were to: 1) provide qualitative evidence to understand meaningful change for patients following clinical intervention and 2) to elicit data on what patients consider to be a minimum meaningful improvement on different patient-reported outcomes (PROs), including:

- The UFS-QoL BPD Scale,
- The PGA symptoms severity.

These objectives were achieved through conducting web/Internet-based video or telephone interviews with English-speaking patients in the US within 3 to 14 days after their Week 24 visit of either ongoing phase 3 clinical study (MVT-601-3001 [LIBERTY 1] or MVT-601-3002 [LIBERTY 2]).

Minimum meaningful improvements on other PROs were also explored as part of the exit interview study; results of the respective exercises will be included in the full report for this exit interview study.

4.4.2 Methodology – Qualitative Interviews

The exit interviews were conducted via a web/Internet-based video platform (Doxy.me [https://doxy.me/]) or via telephone by trained and experienced Endpoint Outcomes interviewers.

In the event that a patient did not improve by at least 1 point from Baseline Day 1 to Week 24 based on her PGA of symptom severity scores, meaningful change exercises were not conducted for any of the PROs. An improvement on the PGA of symptom severity was required so that patients could provide contextually relevant feedback related to positive changes in uterine fibroid symptoms, as they would have experienced an improvement throughout the trial. Table 4.4-1 summarizes the measures/scales of interest, the type of data that was used in the respective meaningful change exercises, and the criteria that must have been met in order for the patient to participate in the respective meaningful change exercise.

Table 4.4-1: Overview of Procedures for Meaningful Change Exercises

Measure/Scale	Type of Data Used	Criteria That Must Have Been Met in Order to Conduct the Respective Meaningful Change Exercise
UFS-QoL BPD Scale (calculated)	MVT-601-3001 (LIBERTY 1) or MVT-601-3002 (LIBERTY 2) Baseline Day 1 response	Improvement on PGA of symptom severity from Baseline Day 1 to Week 24
PGA of symptom severity	MVT-601-3001 (LIBERTY 1) or MVT-601-3002 (LIBERTY 2) responses (Baseline Day 1 and Week 24)	Improvement on PGA of symptom severity from Baseline Day 1 to Week 24

Abbreviations: PGA = patient global assessment; UFS-QoL BPD = Uterine Fibroid Symptom and Health-Related Quality of Life bleeding and pelvic discomfort.

For the UFS-QoL BPD Scale, only patients' clinical study (ie, MVT-601-3001 [LIBERTY 1] or MVT-601-3002 [LIBERTY 2]) Baseline Day 1 data were used during interviews; the meaningful change discussions were hypothetical as Week 24 data were not made available to Endpoint Outcomes.[1] For the UFS-QoL BPD Scale, patients were provided with both their Baseline item-level scores and the summary score calculated based on the three items in the scale. Patients were also given a copy of the three items that comprise the UFS-QoL BPD Scale for reference during the meaningful change exercise. Patients were then presented with prespecified point change increments (ie, 10 points) and asked whether those changes reflected a meaningful improvement. If a patient indicated that a 10-point increment change would be meaningful, she was asked if an increment 5 points fewer would still be meaningful. Using a stepwise approach, interviewers then moved along the scale to identify the point at which minimum meaningful improvement was achieved for the respective patient.

For the PGA of symptom severity, patients were presented with their clinical study scores at Baseline Day 1 and Week 24 and asked if the change was meaningful. Next, patients were presented with a series of hypothetical point changes (ie, more change if the change was not

[1] For secondary endpoint data, only Baseline responses were shared with Endpoint Outcomes.

meaningful or less change if the change was meaningful, as warranted) and asked if those would be meaningful. This process continued until the minimum meaningful change on the PGA of symptom severity for that patient was identified.

Audio recordings of the interviews were transcribed verbatim and anonymized by removing identifying information such as names and places. Each transcript was considered a unit of analysis, and data from all transcripts were aggregated following coding. An initial coding scheme was developed based on the semistructured interview guide and research objectives. The coding scheme was applied and operationalized using Atlas.ti version 8.2.30 (Atlas.ti GmbH, Berlin), a software program designed specifically for qualitative data analysis. Specifically, codes were applied to selected text within each transcript and then queried for frequency across transcripts. Frequencies of patients' interview responses (eg, minimum meaningful change responses) are reported. Minimum meaningful point change medians and ranges were calculated in Excel. As the sample size for the study was small and to reduce the influence of potential outliers, the median is the preferred measure of central tendency reported.

4.4.3 Results

Thirty patients with heavy menstrual bleeding associated with uterine fibroids participated in exit interviews. The average age of these patients was 44, with ages ranging from PPD More than half of the patients (n = PPD self-reported as PPD and most patients (n = PPD were PPD In addition, the majority of patients (n = 26, 86.7%) self-reported some college or higher education as their highest education level. Two patients selected "Other" as the highest level of education and self-reported that they had medical assistant credentials.

The demographic characteristics of the patients from this exit interview study closely matched those of the LIBERTY 1 (MVT-601-3001) and LIBERTY 2 (MVT-601-3002) total sample and the LIBERTY 1 and 2 US sample (see Table 4.4-2). The average age for both the LIBERTY 1 and 2 total sample and US sample was approximately 42 years. Approximately half of participants (n = 396, 51.4%) in the total sample self-reported as black or African American, and over half of the US sample (n = 372, 63.9%) self-reported as black or African American. Additionally, most participants in both the total sample (n = 588, 76.4%) and US sample (n = 450, 77.3%) self-reported as not Hispanic or Latino. Highest level of education data was collected during patient interviews by Endpoint Outcomes; therefore, education level data for all LIBERTY 1 and 2 patients are not available.

Table 4.4-2 includes demographic data for the interviewed study sample as well as the totality of LIBERTY 1 and 2 and the US-based LIBERTY 1 and 2 sample (based on a database snapshot as of 26 Apr 2019).

Table 4.4-2: Patient Demographic Information (from Baseline MVT-601-3001 [LIBERTY 1] or MVT-601-3002 [LIBERTY 2]) and Education Information Collected during Patient Interviews

Baseline Characteristics	Exit Interview Study Sample (N = 30)	LIBERTY 1 and 2 Total Sample (N = 770)	LIBERTY 1 and 2 US Sample (N = 582)
Age (years)			
Mean (SD)	43.9 (4.5)	42.0 (5.4)	42.1 (5.2)
Range	PPD		
Race			
Black or African American	PPD	396 (51.4%)	372 (63.9%)
White		329 (44.4%)	183 (31.4%)
Ethnicity			
Not Hispanic/Latino	PPD	588 (76.4%)	450 (77.3%)
Hispanic/Latino		174 (22.6%)	130 (22.3%)
Highest level of education			
High school (no degree) or less	2 (6.7%)		
High school graduate	2 (6.7%)		
Some college (no degree)	11 (36.7%)		
Associate's degree	4 (13.3%)		
Bachelor's degree	5 (16.7%)		
Master's degree	4 (13.3%)		
Other	2 (6.7%)		

Abbreviations: SD = standard deviation.

Table 4.4-3 below summarizes the total number of exit interview study patients who completed each meaningful change exercise based on the required criteria.

Table 4.4-3: Summary of the Total Number of Exit Interview Study Patients Who Completed Each Meaningful Change Activity

Measure/Scale	Number of Exit Interview Study Patients Participating in Each Exercise (Total N = 30)[2]	Criteria that Must Have Been Met in Order to Conduct the Respective Meaningful Change Exercise
UFS-QoL BPD Scale (calculated)	25	Improvement on PGA of symptom severity from Baseline Day 1 to Week 24
PGA of symptom severity	25	Improvement on PGA of symptom severity from Baseline Day 1 to Week 24

Abbreviations: PGA = patient global assessment; UFS-QoL BPD = Uterine Fibroid Symptom and Health-Related Quality of Life bleeding and pelvic discomfort.

UFS-QoL Bleeding and Pelvic Discomfort Scale

Twenty-five patients improved from Baseline Day 1 to Week 24 on the PGA of symptom severity and participated in the UFS-QoL BPD Scale meaningful change exercise. Data for 24 patients were included in the analysis as one patient provided meaningful change exercise information that was not informative and therefore was excluded from the analysis.[3] The median minimum point change considered to be a meaningful improvement was 10 points (n = 24; range = 5 to 80). The majority of patients completing the UFS-QoL BPD meaningful change activity (n = 15, 62.5%) considered a minimum change of 5 points or 10 points as meaningful (Table 4.4-4).

[2] A total of 30 patients completed exit interviews as part of this study; however, not all 30 patients completed each meaningful change exercise as additional criteria were required in order for a patient to complete the meaningful change exercises. The numbers in this table represent the total number of exit interview patients who met the criteria for participation for the specific meaningful change exercises listed.

[3] This patient did not understand how the three items comprising the UFS-QoL BPD led to the generation of her summary score and could not describe the minimum point change needed for meaningful improvement.

Table 4.4-4: UFS-QoL BPD Scale Meaningful Improvement Results

Minimum Point Change Considered to be a Meaningful Improvement	n (%) [N = 24]
5-point change	11 (45.8%)
10-point change	4 (16.7%)
15-point change	2 (8.3%)
20-point change	0 (0.0%)
25-point change	1 (4.2%)
30-point change	1 (4.2%)
35-point change	1 (4.2%)
40-point change	1 (4.2%)
45-point change	2 (8.3%)
80-point change	1 (4.2%)
Overall point change	
Median	10
Range	5 – 80

Patient Global Assessment of Symptom Severity

Twenty-five patients improved by at least 1 point from Baseline Day 1 to Week 24 on the PGA (for symptoms) and participated in the PGA of symptom severity meaningful change exercise. All patients participating in the PGA of symptom severity meaningful change exercise (n = 25, 100.0%) reported that the actual improvement experienced during the clinical study was meaningful to them.

The median minimum point change considered to be a meaningful improvement was 1 point (n = 24; range = 1 to 3); the most frequently reported minimum meaningful improvement reported by patients was a 1-point change (n = 17, 68.0%) (Table 4.4-5).

Table 4.4-5: PGA Symptom Severity Meaningful Improvement Results

Minimum Point Change Considered to Be a Meaningful Improvement	n (%) [N = 25]
1-point change	17 (68.0%)
2-point change	7 (28.0%)
3-point change	1 (4.0%)
Overall point change	
Median	1
Range	1 – 3

4.4.4 Discussion

The exit interviews provided supportive qualitative evidence to assist in the interpretation of meaningful change in patients following clinical intervention. Patients were required to improve by at least 1 point on the PGA of symptom severity over the course of the clinical study to ensure that patients interviewed had experienced improvement and could reflect upon meaningful improvements in uterine fibroid symptoms.

The decision to use actual clinical trial data in the qualitative interviews was guided by an effort to increase the contextual relevance of each of the meaningful change activities. Providing patients with their Baseline scores for the three PROs created a unique opportunity for patients to reflect on their experience since starting treatment, thereby making the exercises more relevant to them. Further, participation in the meaningful change exercises was predicated on experiencing an improvement in uterine fibroid symptoms over the course of the study, which ensured that patients could speak to meaningful changes stemming from their personal experience. This was confirmed, as all patients participating in the PGA of symptom severity meaningful change exercise (n = 25, 100.0%) reported that the change during the trial was meaningful to them.

These qualitative findings provide patient insight which can be used to supplement psychometric analyses to determine target anchor categories (for the PGA of symptom severity) and responder definitions for the UFS-QoL BPD Scale.

4.5. Determination of Responder Threshold via Triangulation of Findings

Based on the analyses of individual patients' changes in BPD Scale scores, anchored by changes in their response to the PGA of symptom severity, a 20-point change is recommended as the minimum meaningful change threshold for defining a responder. This threshold estimation used the "1-category improvement" PGA group as the target anchor, which is a significantly separated from the "no change" group with respect to the mean change on the BPD Scale. The choice of "1-category improvement" as the target anchor is supported by the majority (17/25, 68%) of the interviewed patients in the exit interview study reporting that a 1-category improvement on the PGA of symptom severity is meaningful to them. The responder threshold of a 20-point change

on the BPD Scale score is larger than what the majority of patients in the exit interview study reported to be meaningful to them, ie, an improvement between 5- to 15-points.

In summary, based on the triangulation of findings from the anchor-based analyses supported by patients' feed-back during exit interviews, a 20-point change in the BPD Scale is proposed as the responder threshold for change in BPD Scale.

4.6. References

Cappelleri JC, Bushmakin AG. Interpretation of patient-reported outcomes. *Statistical methods in medical research* 2014;23:460-483.

Cohen J. Statistical power analysis for the behavioral sciences (1988, 2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

Coon CD, Cook KF. Moving from significance to real-world meaning: methods for interpreting change in clinical outcome assessment scores. *Quality of Life Research* 2018;27:33-40.

Crosby RD, Kolotkin RL, Williams GR. Defining clinically meaningful change in health-related quality of life. *Journal of clinical epidemiology* 2003;56:395-407.Wyrwich KW, Norquist JM, Lenderking WR, Acaster S. Industry Advisory Committee of International Society for Quality of Life R. Methods for interpreting change over time in patient-reported outcome measures. *Qual Life Res* 2013;22:475-483.

Revicki D, Hays RD, Cella D, Sloan J. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *Journal of clinical epidemiology* 2008;61:102-109.

APPENDIX 5. ESTIMATION OF RESPONDER THRESHOLD FOR THE UFS-QOL REVISED ACTIVITIES SCALE

5.1. Approach to Estimating the Responder Threshold of the Revised Activities Scale

The Uterine Fibroid Symptom and Health-Related Quality of Life (UFS-QoL) Revised Activities Scale includes five of the seven most relevant items pertaining to physical and social activities (Coyne 2018). These are:

During the previous 3 months, how often have your symptoms related to uterine fibroids:

- Interfered with your physical activities?
- Made you decrease the amount of time you spent on exercise or other physical activities?
- Made you feel that it was difficult to carry out your usual activities?
- Interfered with your social activities?
- Caused you to plan activities more carefully?

Response options include:

- None of the time;
- A little of the time;
- Some of the time;
- Most of the time;
- All of the time.

The summary score of the five items ranges from 0 to 100, where a lower score indicates a higher ability to do activities (ie, lower score = good) and a higher score indicates a lower ability to do activities.

Change from Baseline to Week 24 in the Revised Activities Scale score is a secondary endpoint of the pivotal studies (MVT-601-3001 and MVT-601-3002) to evaluate the treatment benefit of relugolix + E2/NETA (Group A) compared with placebo (Group C). Additionally, a responder analysis will be performed between the two groups with respect to the proportion of patients who have achieved a meaningful reduction from Baseline to Week 24 in the Revised Activities Scale.

The approach used to derive the responder threshold for improvement in the Revised Activities Scale is similar to that used for the Bleeding and Pelvic Discomfort (BPD) scale (see details in Appendix 4).

This appendix briefly describes the quantitative and supportive qualitative methods and summarizes the respective analysis results.

The meaningful change threshold is the smallest reduction in the Revised Activities Scale score that is considered meaningful by patients (Cohen, 1988; Crosby, 2003; Revicki, 2008; Wyrwich, 2013; Cappelleri, 2014; Coon, 2018). The magnitude of a meaningful change threshold depends

on the magnitude of the correlation between the change in the Revised Activities Scale score and change in anchor (ie, the Patient Global Assessment [PGA] for function anchor) as well as the variability of change on the Revised Activities Scale by improvement categories on the PGA of symptoms (described in Section 5.2.2). Several anchor-based methods will be used; however, the primary analysis will be a measure of central tendency for each improvement category (see Section 5.2.3). Anchor-based methods will use data collected on:

- The UFS-QoL Revised Activities Scale score at Baseline and Week 24; and
- The PGA of function score at Baseline and Week 24.

Results from the anchor-based analyses will be supported by qualitative data collected in a patient interview study (MVT-601-037), a substudy of the phase 3 trials, in which patients from selected sites in the United States (US) provided feedback on what they considered to be a meaningful change on the Revised Activities Scale and the PGA of function (described in Section 5.4).

5.2. Statistical Analysis Plan for Estimation of the Responder Threshold

5.2.1. Anchor and Its Correlation with UFS-QoL Endpoint

The PGA of function uses a five-point verbal rating scale and asks the patient:

> How much were your usual activities limited by uterine fibroid symptoms such as heavy bleeding over the last 4 weeks?

Response options include:

- No limitation at all
- Mild limitation
- Moderate limitation
- Quite a bit of limitation
- Extreme limitation

The categorical change from Baseline to Week 24 in PGA of function score will be derived, leading to nine possible outcomes ranging from +4 (denoting worsening) to -4 (denoting improvement). The change in PGA of function at Week 24 will be used as the anchor (see Table 5.2-1).

5.2.2. Target Anchor Category

The target anchor category is the anchor category that represents the minimum meaningful change and is used as the starting point to identify potential candidates for a meaningful change threshold. For the two pivotal studies, the target anchor category will be a one-point category improvement on the PGA of function (see Table 5.2-1), as this is typically considered as a minimal clinical important difference on a five-point Likert scale.

Table 5.2-1: Change in PGA as Anchor

Anchor	Anchor Change Category	Potential Target Anchor Change Category (To Be Used for Estimation of Meaningful Change Threshold)
Change in PGA of function	-4, -3, -2, -1 (improvement), 0 (same), +1, +2, +3, +4 (worsening)	-1-category change (improvement)

Abbreviations: PGA = patient global assessment.

5.2.3. Anchor-Based Methods

To determine the meaningful change threshold for the reduction in UFS-QoL Revised Activities Scale score, the anchor-based analyses described below will be performed.

The category (or point) change in PGA of function score will be used as the anchor to classify patients into response groups, depending on their level of change in the Revised Activities Scale from Baseline to Week 24 (see Table 5.2-1). Uncollapsed, categorical change on the PGA will range from +4 to -4. Collapsed, categorical change will be considered based on the distribution of change categories on the PGA of function. Usually, the collapsing occurs on the tails with extreme worsening (+4) or improvement (-4).

Among the anchor-based analyses described below, the within-group analysis will be primary and other analyses (including between-group analysis) are supportive.

5.2.3.1. Correlation with Anchor

Correlation between the categorical change on the PGA of function score and the change in the Revised Activities Scale score will be evaluated at Week 24, using blinded pooled data from the first third of the enrolled patients from the two pivotal studies who had completed Week 24 visits and had the corresponding PGA of function data available (denoted as the "threshold determination analysis set"). Polyserial correlation coefficient will be used with a criteria value of > 0.30 indicating meaningful correlation (Cohen, 1988; Crosby, 2003; Revicki, 2008; Cappelleri, 2014; Coon, 2018).

5.2.3.2. Within-Group Meaningful Change

The magnitude of change from Baseline to Week 24 in Revised Activities Scale score will be calculated within each anchor category group. Changes in Revised Activities Scale scores are negative for reduced ability to do activities (indicating a worse outcome) and positive for increased ability to do activities (indicating a better outcome).

Descriptive statistics (*n*, mean change, median change, 25th and 75th percentiles, standard deviation [SD], confidence interval [CI], and standardized effect size [SES]) will be reported for the changes in Revised Activities Scale scores by anchor category. The SES will be calculated for each level of anchor category group by dividing the mean change score of Revised Activities Scale from Baseline by the Baseline SD of the anchor category group. The impact of treatment will be judged based on Cohen's recommendations (1988): small change (SES = 0.20),

moderate change (SES = 0.50), and large change (SES = 0.80). Significance associated within-patient change will be evaluated using paired t-tests on the change in Revised Activities Scale score separately for each level of improvement on the anchor.

5.2.3.3. Supportive Analysis of Between Group Meaningful Change Using Analysis of Variance

Analysis of variance (ANOVA) will be used to determine whether a difference in mean change scores from Baseline to Week 24 on the Revised Activities Scale exists between the categorical change groups (or the collapsed groups, as appropriate). Providing there is a significant change in Revised Activities Scale scores between the (collapsed) anchor groups, the between-group differences will be explored. Any anchor group with at least 15 patients will be included in this analysis. An anchor group with < 15 patients (usually occurring on the tails with extreme worsening [+4] or improvement [-4]) will be collapsed with its adjacent group as appropriate. Comparison of the anchor groups of interest between the target anchor ("-1 change" category) and "0 change" category will be performed using a t-test. A statistically significant difference on the Revised Activities Scale change scores corresponding to a 1-category change on the PGA of function can be used as supportive information for estimating the meaningful change threshold.

5.2.3.4. Visualizing Cumulative Distribution Function and Probability Distribution Function Plots by Anchor Category Group

Anchor-based meaningful change will also be evaluated using cumulative distribution function (CDF) plots utilizing the Kernel smoothing for all anchor category groups, based on cumulative change in the Revised Activities Scale scores for all available changes from Baseline to Week 24. Specifically, the CDF plot for each anchor category displays the probability (presented on y-axis) of patients who have achieved a given absolute change of X or less in the Revised Activities Scale score from Baseline to Week 24 for each point change along the range of possible absolute changes (from -100 [maximum reduction] to 0 [no change] to 100 [maximum increase]) expressed on the x-axis.

Similarly, the smooth probability density function (PDF) will also be plotted for each anchor category group over the range of absolute changes in the Revised Activities Scale scores. These probabilities are plotted on the y-axis with the Revised Activities Scale change score on the x-axis.

The CDF and PDF curves are delineated by anchor improvement category (from -4 to +4) displaying the center and separation between the curve for the target anchor group and the curve for the group reporting no change on PGA of function. It is expected that the CDF curves will not cross between the change category groups (eg, monotonic increase from no change to slightly improved and moderately improved).

5.2.4. Determining a Meaningful Change Threshold Using Totality-of-Evidence Approach

The meaningful change threshold will be determined using the totality of evidence from the results of above quantitative anchor-based analyses; results from the interview study (MVT-601-037) will be used as supportive evidence.

The results of these analyses and proposed thresholds will be included into the Patient-Reported Outcome dossier to be submitted at time of filing.

5.3. Results from Anchor-Based Analyses

5.3.1. Correlation of Change in Revised Activities Scale Score with PGA of Function

Meaningful change for the UFS-QoL Revised Activities Scale was derived based on anchor-based methods, supported by CDF and PDF curves. To assess the suitability of the selected anchor, PGA of function, a polyserial correlation was calculated between change on the PGA from Baseline to Week 24 and the change from Baseline to Week 24 on the Revised Activities Scale. The change in the PGA was moderately negatively correlated (r = -0.60) with the change on the Revised Activities Scale (Table 5.3-1). Given that the PGA of function is less complex than the Revised Activities Scale, this result indicates that the PGA of function is a suitable anchor for the Revised Activities Scale.

Table 5.3-1: Summary of Change from Baseline to Week 24 in UFS-QoL Revised Activities Scale by PGA of Function Change Category (mITT Population)

PGA of Function Change Category	N = 254	Change in Revised Activities					Correlation between PGA Change and Revised Activities Change[a]
		Mean (SD)	Median	95% CI	p-value[b]	SES[c]	
4-category deterioration (+4)	**2**	5.00 (7.07)	5	-58.53,68.53	0.500	0.28	-0.60
3-category deteriorations (+3)	**2**	0	0	-	-	0.00	
2-category deteriorations (+2)	5	7.00 (22.80)	0	-21.31,35.31	0.5302	0.61	
1-category deteriorations (+1)	**22**	-1.59 (23.82)	-5	-12.15,8.97	0.7572	-0.06	
0 Category deteriorations (0)	71	11.55 (28.51)	5	4.80,**18.30**	0.0011	0.38	
1-category improvement (-1)	53	**27.92** (25.65)	**20**	**20.85**,35.00	< 0.0001	1.06	
2-category improvement (-2)	**51**	51.86 (27.60)	60	44.10,59.63	< 0.0001	2.17	
3-category improvement (-3)	**35**	56.81 (27.49)	57.50	47.50,66.11	< 0.0001	2.91	
4-category improvement (-4)	13	60.77 (31.55)	70	41.71, 79.83	< 0.0001	4.40	

Abbreviations: CI = confidence interval; mITT = modified intent-to-treat; PGA = patient global assessment; SD = standard deviation; SES = standardized effect size.
mITT is used to calculate change from Baseline score at Week 24 and includes patients from the mITT population who have available change from Baseline data at Week 24.

[a] Polyserial correlation coefficient between change in Revised Activities Scale and change in PGA of function.

[b] The p-value for each individual change group is derived from a paired (within-sample) t-test assessing the difference over time.

[c] SES calculated as the mean divided by the SD of Baseline. SES is judged as small = 0.2, moderate = 0.5, and large = 0.8 (Cohen 1988).

5.3.2. Improvement on Revised Activities Scale by PGA Change Category

Uncollapsed changes on the PGA of function were used to determine minimal meaningful improvement on the Revised Activities Scale (Table 5.3-1). Improvement on the Revised Activities Scale increased monotonically for all the categories from "no change (0)" to "1-category improvement (-1)" to "2-category improvement (-2)" with non-overlapping 95% CIs for mean change of the three groups. Table 5.3-2 shows that a one category improvement (-1) is associated with a 27.92-point mean improvement in the Revised Activities Scale score at Week 24 compared to Baseline, with a 95% CI [20.85, 35.00], a large SES = 1.06, and a median improvement of 20 points.

Table 5.3-2 highlights that the difference between the "1-category improvement" and the "no change" groups (mean =11.55 with a 95% CI of [4.80, 18.30]) was statistically significant (p = 0.0013) with a moderate SES = 0.54, which reasonably supports the notion that patients interpreted these change categories as distinct.

Table 5.3-2: Summary of Change from Baseline to Week 24 in Revised Activities Scale Between Target Anchor (-1) and No change (0) in PGA of Function (mITT Population)

Anchor	Categorical Change	N	Mean Change from BL	SD	95% CI	p-value[a]	Baseline SD	SES
PGA	1-category improvement (-1)	53	**27.92**	**25.65**	**20.85, 35.0**			
	No change (0)	71	**11.55**	**28.51**	**4.80, 18.30**			
	Difference		**16.38**	**27.33**	**6.55, 26.20**	0.0013		0.54[b] 0.57[c]

[a] The p-value is based on t-test for difference in mean change in BPD score between the 2 anchor groups (-1 and 0) from the ANOVA.

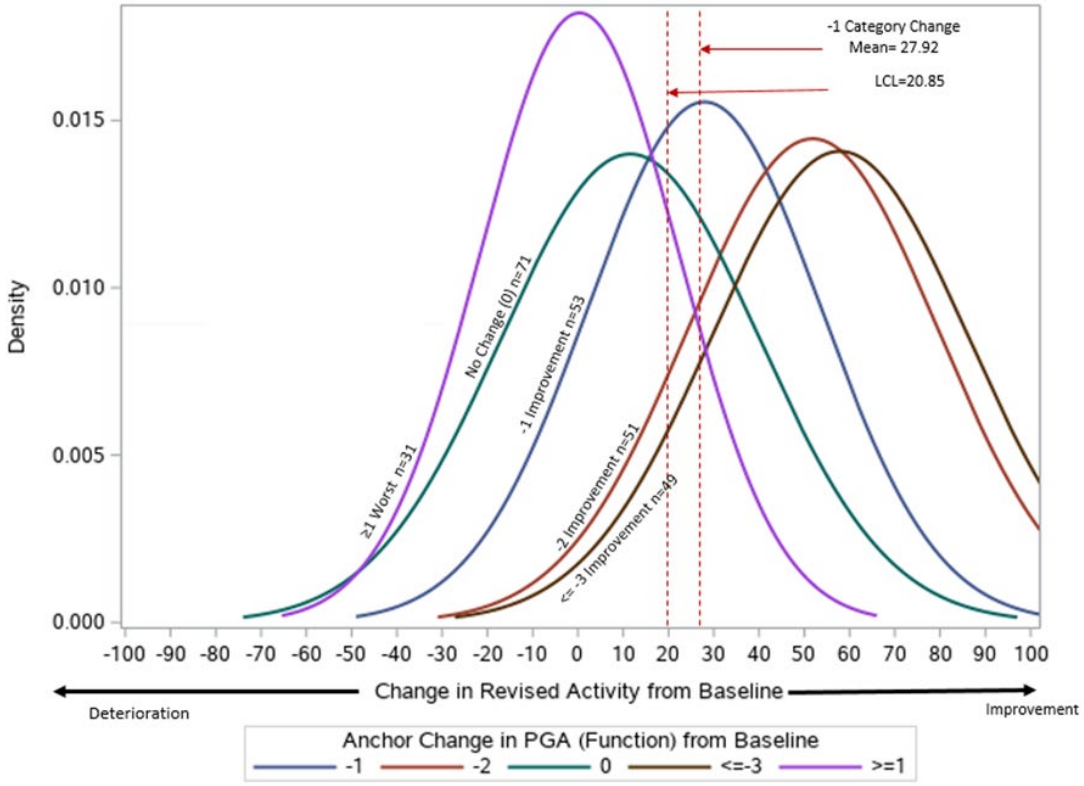
[b] SES calculated as the mean difference divided by the standard deviation of Baseline for no change group. They are judged as small=0.2, moderate=0.5 and large=0.8 (Cohen, 1988).

[c] SES calculated as the mean difference divided by the standard deviation of Baseline for pooled from all categories (Glass, 1976).

That patients were able to distinguish between the PGA "1-category improvement" and the 'no change" group is further supported by the nonoverlapping CIs (in Table 5.3-2) for the respective UFS-QoL Revised Activities Scale scores and as illustrated by the separation between the CDF curves presented in Figure 5.3-1. Since statistically significant differences existed in patient responses on the Revised Activities Scale between the "1-category improvement (-1) " option and the "no change" and the "2-category improvement (-2)" groups, a 1-category improvement on the PGA was considered a meaningful target anchor category for assessing the responder threshold on the Revised Activities Scale. Although a two-category improvement could have been considered for deriving the meaningful change threshold, such a threshold would not qualify as being the *minimum* threshold possible. The evidence (ie, the statistical difference between the1- and 2-category improvements and the fact that patients were able to distinguish between the two response options) supports using a 1-category improvement on the PGA of

function for estimating the minimum meaningful change threshold. This decision is also supported by qualitative evidence generated from the Exit Interview study (see Section 5.4).

Figure 5.3-1: PDF of the Change in UFS-QoL Revised Activities by PGA of Function Anchor Change Category (Collapsed)



Abbreviations: PGA = patient global assessment; LCL = lower confidence limit.

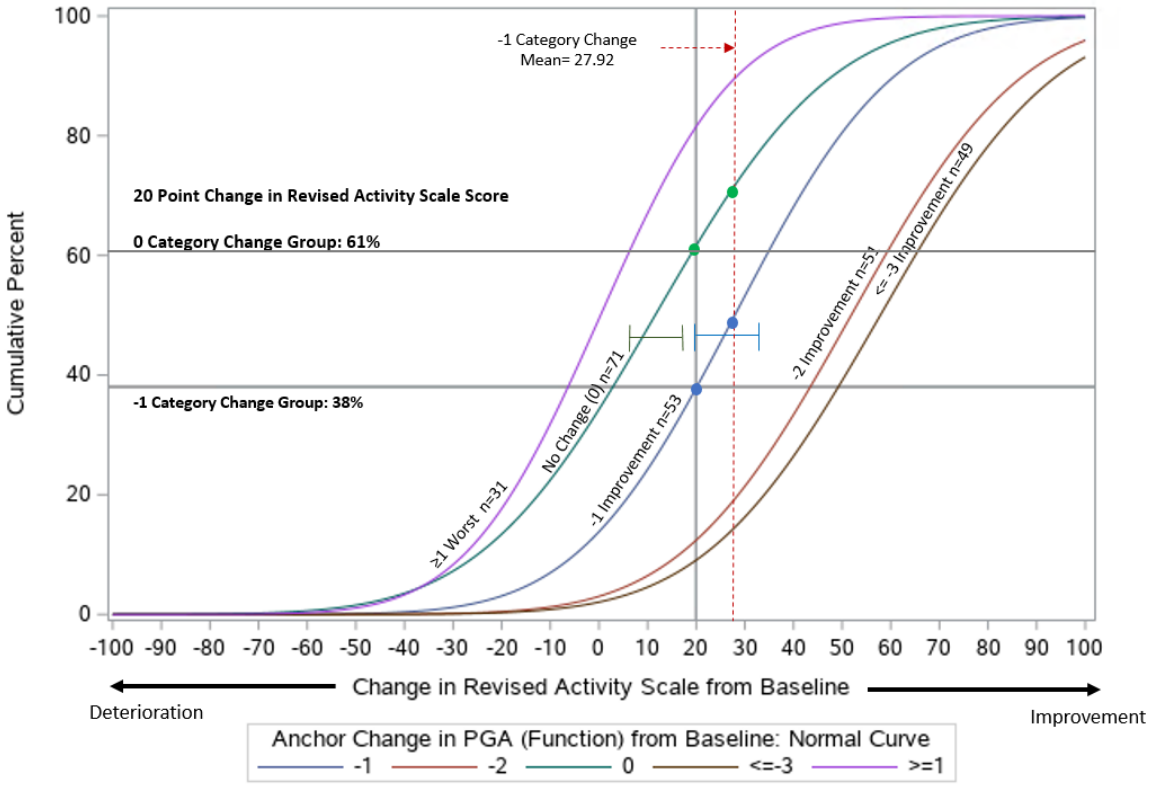
5.3.3. Estimation of Responder Threshold

Using the mean value for measuring improvement in the Revised Activities Scale would yield estimates that are conservative because expected values do not necessarily constitute a *minimum* meaningful change threshold for patients. That is, nearly half the patients stratified in the PGA "1-category improvement" who reported changes smaller than the mean on the Revised Activities Scale would be classified as nonresponders by using the mean as the threshold despite of their reporting "1-category improvement". A less conservative, though still plausible estimate for the minimal meaningful change threshold is the lower bound of the 95% CI for mean change in the "1-category improvement" group. Its use will result in a smaller proportion of patients being classified as nonresponders on the Revised Activities Scale than the expected value (ie, the mean). Similarly, one can also consider the median value since it is less influenced by outliers than either the mean or CI estimates.

According to the uncollapsed anchor-based analysis (Table 5.3-1), the median value for a "1-category improvement" is 20-points, while the lower bound 95% CI for this group is about 21-points (ie, a 21-point improvement on the revised activities between Baseline and Week 24). Given the large discrepancy between the mean and median values suggests that outliers were present in the data; hence, the median value is recommended as a potential minimum change threshold.

Examination of the CDF curves for the potential minimum meaningful threshold value of 20 points on the Revised Activities Scale allows one to estimate the cumulative percent of patients that would experience the improvement. As illustrated in Figure 5.3-2, approximately 38% of the "no change" group and 61% of the "1-category improvement" group experienced at least a 20-point improvement (eg, approximately 62% of the "no change" group and 39% of the "1-category improvement" group experienced less than a 20-point improvement to the left) on the Revised Activities Scale by Week 24.

Figure 5.3-2: Cumulative Distribution Function of Change at Week 24in UFS-QoL Revised Activities Scale Score by PGA Anchor Change Category (Collapsed)

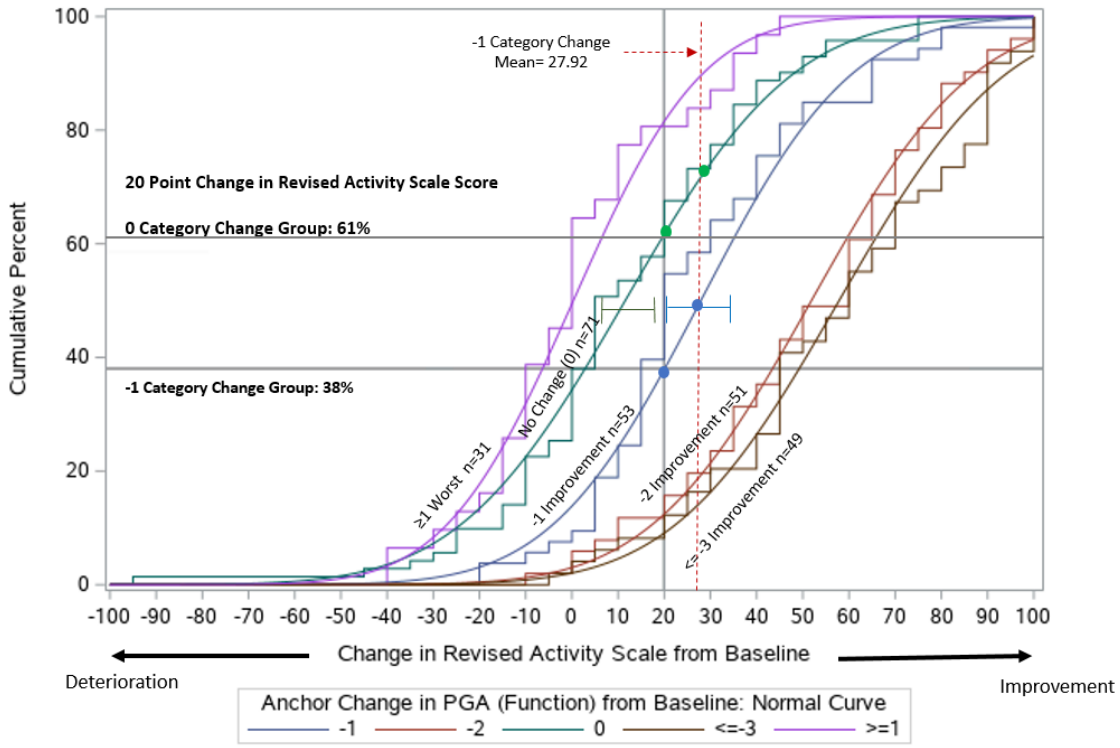


Abbreviations: PGA = patient global assessment.

As supportive information, the empirical CDFs with step-curves (reflecting the discrete nature of the revised activities scores) are provided (Figure 5.3-3), indicating that smooth curves are

reasonably close to the empirical CDFs. Examination of the PDF curves presented in Figure 5.3-1 indicates that the dispersion is roughly the same for the options between "> -3-category improvement" and "no change."

Figure 5.3-3: Empirical Cumulative Distribution Function of Change at Week 24 in UFS-QoL Revised Activities Scale Score by PGA Anchor Change Category (Collapsed)



Abbreviations: PGA = patient global assessment.

5.4. Exit Interview Study Synthesis

5.4.1 Objectives

The objectives of the exit interviews were: 1) to provide qualitative evidence to understand meaningful change for patients following clinical intervention and 2) to elicit data on what patients consider to be a minimum meaningful improvement on different patient-reported outcomes (PROs), including:

- The UFS-QoL Revised Activities Scale;
- The PGA of function.

These objectives were achieved through conducting web/Internet-based video or telephone interviews with English-speaking patients in the US within 3 to 14 days after their Week 24 visit

of either ongoing phase 3 clinical study (MVT-601-3001 [LIBERTY 1] or MVT-601-3002 [LIBERTY 2]).

Minimum meaningful improvements on other PROs were also explored as part of the exit interview study; results of the respective exercises will be included in the full report for this exit interview study.

5.4.2 Methodology – Qualitative Interviews

The exit interviews were conducted via a web/Internet-based video platform (Doxy.me [https://doxy.me/]) or via telephone by trained and experienced Endpoint Outcomes interviewers.

If a patient did not improve by at least 1 point from Baseline Day 1 to Week 24 based on her PGA of function, meaningful change exercises were not conducted for the PGA of function and the UFS-QoL Revised Activities Scale. An improvement on the PGA of function was required so that patients could provide contextually relevant feedback related to positive changes as they would have experienced an improvement throughout the trial. Table 5.4-1 summarizes the measures/scales of interest, the type of data that was used in the respective meaningful change exercises, and the criteria that must have been met in order for the patient to participate in the respective meaningful change exercise.

Table 5.4-1: Overview of Procedures for Meaningful Change Exercises

Measure/Scale	Type of Data Used	Criteria That Must Have Been Met in Order to Conduct the Respective Meaningful Change Exercise
UFS-QoL Revised Activities Scale (calculated)	MVT-601-3001 (LIBERTY 1) or MVT-601-3002 (LIBERTY 2) Baseline Day 1 response	Improvement on PGA of function from Baseline Day 1 to Week 24
PGA (for function)	MVT-601-3001 (LIBERTY 1) or MVT-601-3002 (LIBERTY 2) responses (Baseline Day 1 and Week 24)	Improvement on PGA of function from Baseline Day 1 to Week 24

Abbreviations: PGA = patient global assessment; UFS-QoL = Uterine Fibroid Symptom and Health-Related Quality of Life.

For the UFS-QOL Revised Activities Scale, only patients' clinical study (ie, MVT-601-3001 [LIBERTY 1] or MVT-601-3002 [LIBERTY 2]) Baseline Day 1 data were used during interviews; the meaningful change discussions were hypothetical, as Week 24 data were not made available to Endpoint Outcomes.[4] For the UFS-QoL Revised Activities Scale, patients were provided with both their Baseline item-level scores and the summary score calculated based on the five items in the scale. Patients were also given a copy of the five items that comprise the UFS-QoL Revised Activities Scale for reference during the meaningful change exercise. Patients were then presented with pre-specified point change increments (ie, 10 points) and asked whether those changes reflected a meaningful improvement. If a patient indicated that a 10-point

[4] For secondary endpoint data, only Baseline responses were shared with Endpoint Outcomes.

increment change would be meaningful, she was asked if an increment 5 points fewer would still be meaningful. Using a stepwise approach, interviewers then moved along the scale to identify the point at which minimum meaningful improvement was achieved for the respective patient.

For the PGA of function, patients were presented with their clinical study scores at Baseline Day 1 and Week 24 and were asked if the change was meaningful. Next, patients were presented with a series of hypothetical point changes (ie, more change if the change was not meaningful or less change if the change was meaningful, as warranted) and asked if those would be meaningful. This process continued until the minimum meaningful change on the PGA of function for that patient was identified.

Audio-recordings of the interviews were transcribed verbatim and anonymized by removing identifying information such as names and places. Each transcript was considered a unit of analysis, and data from all transcripts were aggregated following coding. An initial coding scheme was developed based on the semi-structured interview guide and research objectives. The coding scheme was applied and operationalized using Atlas.ti version 8.2.30 (Atlas.ti GmbH, Berlin), a software program designed specifically for qualitative data analysis. Specifically, codes were applied to selected text within each transcript and then queried for frequency across transcripts. Frequencies of patients' interview responses (eg, minimum meaningful change responses) are reported. Minimum meaningful point change medians and ranges were calculated in Excel. As the sample size for the study was small and to reduce the influence of potential outliers, the median is the preferred measure of central tendency reported.

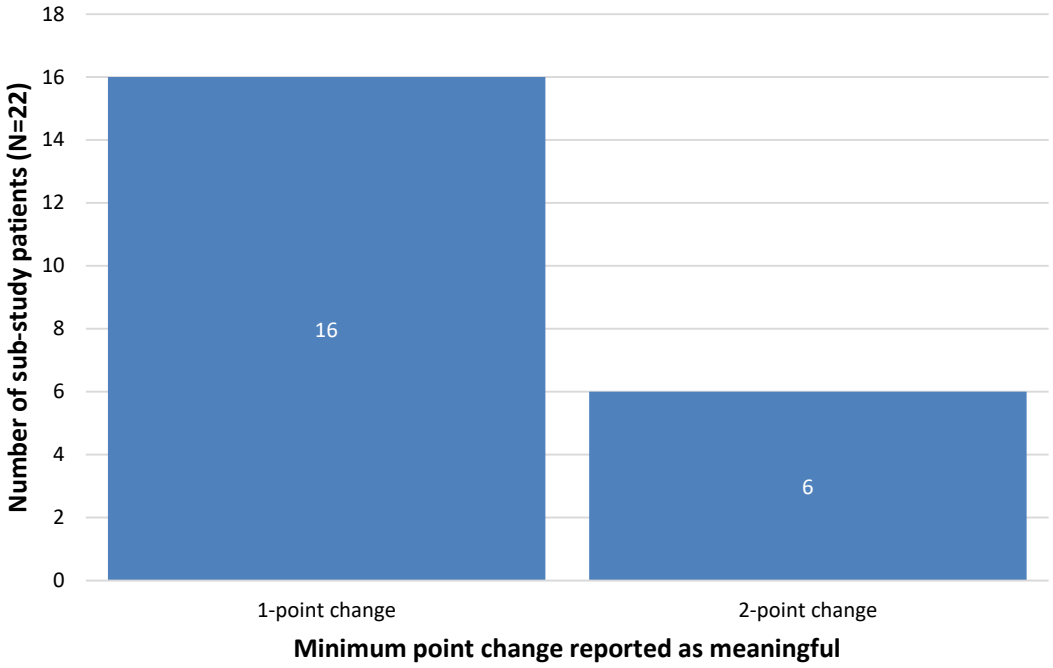
5.4.3 Results

5.4.3.1 PGA of Function[5]

Twenty-two patients improved from Baseline Day 1 to Week 24 on the PGA of function and participated in the PGA of function meaningful change exercise. The demographic characteristics of the 22 patients who completed the PGA of function closely match that of the entire substudy sample as the sample was mostly PPD (n = PPD (n = PPD had completed at least some college or higher (n = 19, 86.4%), and had an average age of approximately 44 years. The median minimum point change considered to be a meaningful improvement was 1 point (n = 22, range = 1-2); the most frequently reported minimum meaningful improvement reported by patients was a 1-point change (n = 16, 72.7%) (Figure 5.4-1).

[5] The PGA of function asks: How much were your usual activities limited by uterine fibroids symptoms such as heavy bleeding over the last 4 weeks? Response options include: No limitation at all, mild limitation, moderate limitation, quite a bit of limitation, and extreme limitation.

Figure 5.4-1: Meaningful Change Estimation: Results of the PGA (for Function)



5.4.3.1 UFS-QoL Revised Activities Subscale[6]

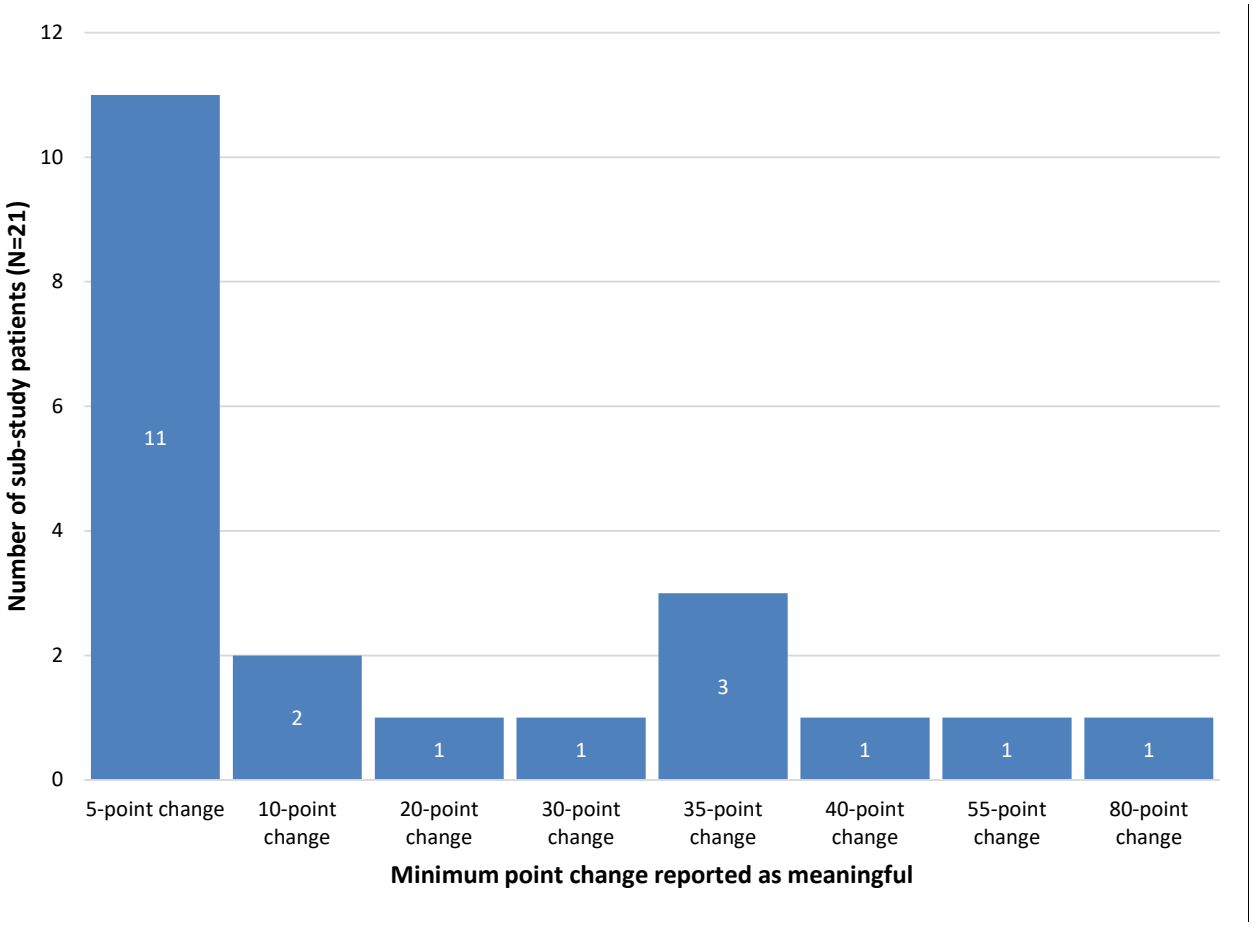
Twenty-two patients improved from Baseline Day 1 to Week 24 on the PGA of function and participated in the UFS-QoL revised activities subscale meaningful change exercise. Data for 21 patients were included in the analysis as one patient provided meaningful change exercise information that was not informative and therefore was excluded from the analysis.[7] The demographic characteristics of the 21 patients who completed the UFS-QoL Revised Activities Scale closely match that of the entire substudy sample as the sample was mostly PPD (n = PPD (n = PPD had completed at least some college or higher (n = 19, 90.5%), and had an average age of approximately 44 years.

[6] The UFS-QoL revised activities subscale includes five items, which ask: During the previous 3 months, how often have your symptoms related to uterine fibroids … 11) interfered with your physical activities; 13) made you decrease the amount of time you spent on exercise or other physical activities; 19) made you feel it was difficult to carry out your usual activities; 20) interfered with your social activities; and 27) made you plan activities more carefully. Response options include 1) None of the time, 2) A little of the time, 3) Some of the time, 4) Most of the time, and 5) All of the time. The score range for the subscale is 0-100. A higher score on the revised activities subscale indicates a lower interference in activities while a lower score on the subscale indicates a higher interference in activities.

[7] This patient was unwilling to describe the minimum point change needed for meaningful improvement for the UFS-QoL revised activity subscale.

The median minimum point change considered to be a meaningful improvement was 5 points (n = 21, range = 5-80); the most frequently reported minimum meaningful improvement reported by patients was a 5-point change (n = 11, 52.4%) (Figure 5.4-2).

Figure 5.4-2: Meaningful Change Estimation: Results of the UFS-QoL Revised Activities Subscale



5.5. Determination of Responder Threshold via Triangulation of Findings

Based on the analyses of individual patient's change in Revised Activities Scale scores anchored by change in their response to the PGA of function, a 20-point change is recommended as the minimum meaningful change threshold for defining a responder. This threshold estimation used the "1-category improvement" PGA group as the target anchor, which is significantly separated from the "no change" group with respect to the mean change on the Revised Activities Scale. The choice of "1-category improvement" as the target anchor is supported by the majority (16/22, 73%) of the interviewed patients in the exit interview study reporting that a 1-category improvement on the PGA of function is meaningful to them. The responder threshold of a 20-point change on the Revised Activities Scale score is larger than what the majority of patients in the exit interview study reported to be meaningful to them (ie, improvements of 5 points [11/21] and 10 points [2/21]).

In summary, based on the triangulation of findings from the anchor-based analyses supported by patients' feedback during exit interviews, a 20-point change in the Revised Activities Scale is proposed as the responder threshold for change in Revised Activities Scale.

5.6. References

Cappelleri JC, Bushmakin AG. Interpretation of patient-reported outcomes. Stat Meth Med Res 2014;23:460-483.

Cohen J. Statistical power analysis for the behavioral sciences (1988, 2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

Coon CD, Cook KF. Moving from significance to real-world meaning: methods for interpreting change in clinical outcome assessment scores. Qual Life Res 2018;27:33-40.

Crosby RD, Kolotkin RL, Williams GR. Defining clinically meaningful change in health-related quality of life. J Clin Epidemiol 2003;56:395-407.

Coyne KS, Harrington A, et al. Psychometric validating of the 1-month recall Uterine Fibroid Symptom and Health-Related Quality of Life Questionnaire (UFS-QOL). ISPOR 23rd Annual International Meeting, 2018

Revicki D, Hays RD, Cella D, Sloan J. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. J Clin Epidemiol 2008;61:102-109.

Wyrwich KW, Norquist JM, Lenderking WR, Acaster S. Industry Advisory Committee of International Society for Quality of Life R. Methods for interpreting change over time in patient-reported outcome measures. Qual Life Res 2013;22:475-483.