Document Type:	Study Analysis Plan
Official Title:	A randomized, parallel-group, double-blind and placebo-controlled, multicenter study to assess the efficacy and safety of vilaprisan in subjects with uterine fibroids
NCT Number:	NCT03400956
Document Date:	20 MAY 2020

Title page

A randomized, parallel-group, double-blind and placebo-controlled, multicenter study to assess the efficacy and safety of Vilaprisan in subjects with uterine fibroids

Assess **S**afety and Efficacy of Vilaprisan in Subjects with U**ter**ine Fibr**oid**s

Acronym: ASTEROID 4

Bayer study drug	BAY 1002670 / Vilaprisan		
Study purpose:	Assessment of safety and efficacy		
Clinical study phase:	III	**Date:**	20 MAY 2020
Study No.:	15790	**Version:**	1.0
Author:	PPD		

This Statistical Analysis Plan is produced on a word-processing system and bears no signatures.

The approval of the Statistical Analysis Plan is documented in a separate Signature Document.

Table of Contents

Table of Tables

Table of Figures

Abbreviations

AE	Adverse event
AESI	Adverse event of special interest
AH	Alkalin hematin
ALT	Alanine aminotransferase (also known as GPT)
AP	Alkaline phosphatase
AST	Aspartate aminotransferase (also known as GOT)
ATC	Anatomical Therapeutic Chemical
BMI	Body Mass Index
BMQ	Bayer MedDRA query
BR	Treatment break
BRM	Blind review meeting
CGI-I	Clinician Global Impression Investigator
CMH	Continuity-corrected Cochran-Mantel-Haenszel
DSMB	Data safety monitoring board
E2	Estradiol
eCRF	Electronic case report form
eDiary	Electronic diary
EIN	Endometrioid Intraepithelial Neoplasia
ePRO	Electronic patient-reported outcome
EoT	End of treatment
FAS	Full analysis set
FSH	Follicle-stimulating hormone
FUP	Follow-up period
HbA1c	Glycated hemoglobin
HDL	High density lipoprotein
HMB	Heavy menstrual bleeding
HRQoL	Health Related Quality of Life
LH	Luteinizing hormone
LDL	Low density lipoprotein
LKF	Laboratorium für Klinische Forschung
LLOQ	Lower limit of quantification
MAR	Missing at random
MNAR	Missing not at random
MedDRA	Medical Dictionary for Regulatory Activities
MBL	Menstrual blood loss
MP	Menstrual pictogram
M&S plan	Modelling and Simulation plan
NRC	National Research Council
P	Progesterone
PAEC	Progesterone receptor modulator-associated endometrial changes
PGI-C	Patient Global Impression of Change
PGI-S	Patient Global Impression of Severity
pH	Potential of hydrogen
PPS	Per-protocol set
PRO	Patient-reported outcome
RAVE	Electronic data capturing system
RND	Randomization
SAE	Serious adverse event
SAF	Safety analysis set
SAP	Statistical Analysis Plan
SAS	Statistical Analysis Software
SD	Standard deviation
SOC	System organ class
TEAE	Treatment-emergent adverse event
TPx	Treatment Period x, $x = 1,2$
TSH	Thyroid-stimulating hormone
UF-DBD	Uterine Fibroid Daily Bleeding Diary

UF-DSD	Uterine Fibroid Daily Symptom Diary
UFS-QoL	Uterine Fibroid Symptom and Quality of Life Questionnaire
ULN	Upper limit of normal
ULOQ	Upper limit of quantification
WHO (-DD)	World Health Organization (-Drug Dictionary)

1. Introduction

Uterine fibroids are benign tumors originating from smooth muscle cells of the myometrium. The pathophysiology of fibroids is not well understood. Genetic predisposition, exposure to steroid hormones, and growth factors play a role in formation and growth.

Uterine fibroids typically appear and grow during reproductive years, but stabilize or regress after menopause. Therefore, they rarely require treatment after menopause. Clinically, fibroids and associated symptoms are most prominent in the late reproductive years. The most common symptoms of uterine fibroids are heavy menstrual bleeding (HMB) and pelvic discomfort.

Uterine fibroids are the leading cause for hysterectomy. Hysterectomy is the only definitive treatment so far and eliminates the possibility of recurrence. The hysterectomy rate has decreased recently but still accounts for almost three quarters of all fibroid related surgical procedures. Increasingly more women desire to avoid hysterectomy, electing for a uterine preserving procedure, regardless of whether they desire to retain their fertility. The surgical treatment options are numerous, and each carries both the risks for surgery itself, as well as the possibility that the woman may require subsequent surgery as new fibroids often develop over time and become symptomatic.

The development of selective progesterone receptor modulators offers the potential for a novel, well tolerated medical treatment approach for women who are experiencing symptoms caused by their fibroids. Various studies have demonstrated the steroid-dependence of fibroid growth and that progesterone has a critical role.

This clinical study is part of a development program for vilaprisan. The rationale of the study is to assess the efficacy and the safety of vilaprisan in subjects with uterine fibroids.

The statistical analysis plan (SAP) is based on the following study protocol documents:

- Integrated Clinical Study Protocol, version 6.0, dated 17 FEB 2020
- Integrated Clinical Study Protocol, version 5.0, dated 21 NOV 2019
- Global Protocol Amendment, version 4.0, dated 11 DEC 2018
- Integrated Clinical Study Protocol, version 3.0, dated 04 JUL 2018
- Integrated Clinical Study Protocol, version 2.0, dated 15 NOV 2017
- Clinical Study Protocol, version 1.0, dated 01 SEP 2017

2. Study Objectives

The primary objective of this study is to show superiority in treatment of HMB of vilaprisan in subjects with uterine fibroids compared to placebo.

The secondary objectives of this study are to additionally evaluate the efficacy and safety of vilaprisan in subjects with uterine fibroids. With the implementation of Integrated Clinical Study Protocol, version 5.0, additional focus will be put on safety evaluations of the endometrium, adrenal glands and skin.

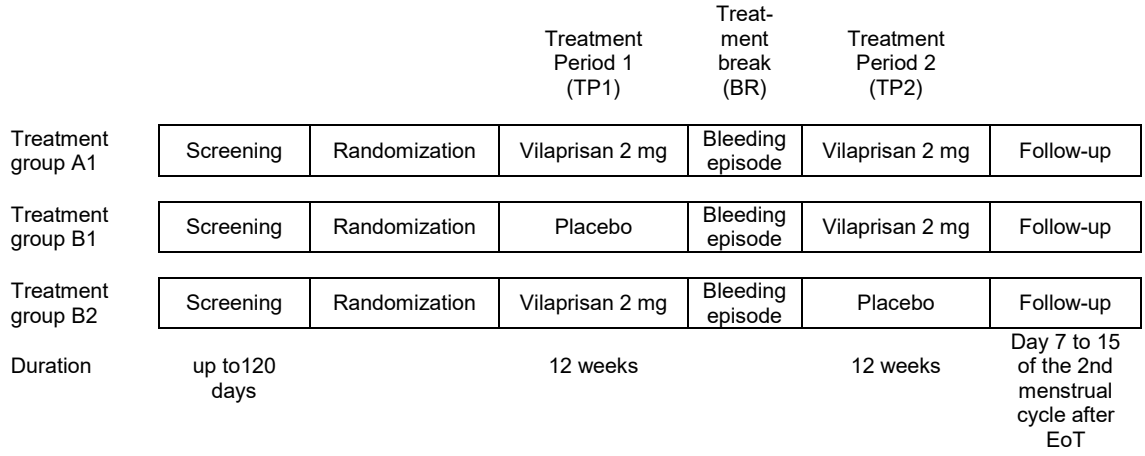
The other objectives of this study are to evaluate the variability in exposure in relation to the efficacy and safety for vilaprisan and to collect patient-reported outcome (PRO) and clinician-reported outcome data.

3. Study Design

This is a randomized, parallel-group, double-blind, placebo-controlled, multicenter study. The study was conducted in the US, Japan and other countries in Europe.

An overview of the study design before the temporary pause is shown in Figure 3–1.

Figure 3–1: Study design as originally planned prior to Global Protocol Amendment, version 4.0



Abbreviations: EoT: End of treatment

The following procedures were implemented prior to the Global Protocol Amendment, version 4.0:

- During the screening period, subjects needed to demonstrate eligibility including the presence of at least 1 uterine fibroid ≥ 30 mm and < 120 mm in largest diameter based on ultrasound, HMB in at least 2 bleeding periods, each with menstrual blood loss (MBL) > 80.00 mL documented by the alkaline hematin (AH) method, and endometrial biopsy results without significant histological disorder.
- Eligible subjects were randomized in 1:1:1 ratio to one of the three treatment groups (A1, B1, or B2) and stratified by country/region (US, Japan and other countries). The time between randomization and start of treatment should not have exceeded 40 days.
- After the end of the final treatment period, subjects were followed up until Day 7 to Day 15 of the 2nd menstrual cycle after end of treatment.

With the implementation of the Global Protocol Amendment, version 4.0, no further subjects were recruited. Patients who were randomized but had not yet started treatment were instructed to not start TP1. Patients already in a treatment period were instructed to complete the current treatment period but to not start a new treatment period. Thus, with the implementation of the Global Protocol Amendment, version 4.0, the originally foreseen study

design outlined in Figure 3–1 is not applicable anymore. With the Integrated Clinical Study Protocol version 5.0 the confirmatory analysis was modified accordingly: The confirmatory analysis (originally including TP1 and TP2) will only be based on results of TP1 as the number of patients who start TP2 is not sufficient for drawing valid conclusions.

All subjects who were randomized and started treatment before the temporary pause were asked to have a comprehensive safety evaluation (with particular focus on endometrial, adrenal and skin safety) performed. This also applies to subjects who have completed or discontinued the study before or during the temporary pause, provided they have taken at least one dose of study medication.

4. General Statistical Considerations

4.1 General Principles

The statistical analysis will be performed using the software package SAS release 9.4 (SAS Institute Inc., Cary, NC, USA).

The following Bayer standards are applied: RD-OI-0119 ‘Prepare statistical documents and programs’, RD-SOP-1107 ‘Recording and Evaluation of Bleeding Data’ together with corresponding Bayer standard SAS macro BCycle (version 6.0). Furthermore, global standards for tables and data displays are available and will be specified in the tables, listings and figures (TLF) specification document, which will be approved in parallel with this SAP.

All variables will be analyzed by descriptive statistical methods according to their data type. The number of data available, mean, standard deviation (SD), minimum, median and maximum will be calculated for metric data as appropriate. The number of significant digits for minimum and maximum will correspond to that of the original data. The mean and median will be rounded to one more decimal place than the original data, and the SD it will be rounded to two more decimal places than the original data. Frequency tables will be generated for categorical data including percentages with 1 decimal place. In general, the data will be presented by treatment group (see Figure 3–1) or by treatment period within treatment group, and overall, that is, based on the total study population of the respective analysis set. Tables, figures and listings will show the treatment groups, A1, B1 and B2. Additionally, for selected variables the treatment of vilaprisan in TP1 of the treatment groups A1 and B2 will be presented also as pooled.

Variables recorded in the electronic case report form (eCRF) and relevant derived variables will be shown in subject data listings, whereby only randomized patients will be included. Data from screening failures will only be shown in the ‘Screening failure’ listing.

Analyses will be performed based on either visits or time intervals as specified in the respective sections. For visit-based analyses the EoT visit will include also (scheduled or unscheduled) EoT visits for subjects who did not start TP2 for any reason, for example, due to the temporary pause of the studies or due to premature discontinuation at the subject’s wish. For time interval-based analyses of efficacy variables (except for fibroid surgeries), including analyses for the treatment break and FUP, analyses for the FUP will be based on data from patients who have started TP2. Data from FUP for patients who did not start TP2 for any reason will be included in the analysis of the break period. For those patients only efficacy-related data up to a theoretical end of the break will be considered (see ‘break period’ in Section 4.5 for details). In contrast to that, for time interval-based analyses of safety variables

(except for bleeding patterns per 28 or 84 days based on UF-DBD), data from FUP for patient who did not start TP2 for any reason will be included in the analysis of the FUP.

4.2 Handling of Dropouts

A subject who has been randomized and discontinues study participation prematurely for any reason is defined to be a *dropout* even if no study drug has been taken. Dropouts will not be replaced.

The number of subjects, who prematurely discontinue the study, as well as the reasons for premature discontinuations, will be reported.

The handling of missing data is described in Section 4.3.

4.3 Handling of Missing Data

Missing data introduce ambiguity into the analysis, beyond the familiar sampling imprecision. Even with the best planning and despite best efforts, data may be missing at the end of the clinical trial. All missing or partial data will be presented in the subject data listing as they are recorded in the eCRF, electronic diary (eDiary) or tablet computer.

When appropriate and for a limited number of variables, specific rules will be implemented so as not to exclude subjects or observations from statistical analyses due to missing or incomplete data. The rules are outlined in the following subsections.

4.3.1 Bleeding data

The following rules for the handling of missing vaginal bleeding data pertains to bleeding data assessed by the AH method, the menstrual pictogram (MP, see Appendix 1) or the item "Rate the severity of **any vaginal bleeding** in the past 24 hours" (on a 6-point Likert scale) in the Uterine Fibroid Daily Bleeding Diary (UF-DBD, see Appendix 2). The resulting data forms the basis for the primary, the secondary and bleeding-related other efficacy variables (see Section 6.2).

On days without bleeding, no collection of sanitary products for the AH value determination was asked for. The UF-DBD is used for the identification of these non-bleeding days. Thus, missing AH/MP values on days where the UF-DBD recorded the bleeding intensity as "No vaginal bleeding" or "Spotting" will be set to 0 mL.

Missing bleeding data in the subjects' diaries will be imputed. First, missing bleeding intensities in the UF-DBD will be imputed. Missing bleeding intensities will only be imputed if values are missing for a single or for two consecutive days. For the UF-DBD, a type of 'worst case approach' will be used for imputing missing bleeding intensities. The approach consists in replacing the missing value(s) by the maximum of the bleeding intensities of the day before and the day after the day(s) with missing value(s).

After having imputed missing values in the UF-DBD, missing AH/MP values are imputed in the next step.

For days with bleeding intensity of 'mild' or higher in the UF-DBD, missing AH/MP values will be replaced by the mean value of AH/MP values of the days with the same bleeding intensity. In case there are no AH/MP values with the same bleeding intensity, the mean of the AH/MP values of next higher intensity will be used. In case there is no intensity higher than the bleeding intensity of the missing AH/MP values available, no replacement for the missing AH/MP values will be done for such days, but if at least one such day occurs within the respective 28 days of treatment for a subject, the subject will be considered as not having amenorrhea.

In case blood loss could not be measured by the vendor for a sanitary product, the missing value will be replaced by the mean blood loss measured for the same type of sanitary product for this specific subject.

4.3.2 Dates for MBL assessed by AH method

If there is no information on the dates of reported AH values, the AH values will not be discarded from the analysis. Those AH values will be attributed to the subject's last bleeding episode prior to the date of receipt of the sample at vendor "Laboratorium für Klinische Forschung" (LKF). For this purpose, the total MBL for each date of receipt will be derived using the AH values with missing date. The total MBL will then be added to equal amounts to the AH values of the subject's last bleeding episode prior to receipt of the sample at LKF.

The concrete steps are outlined below:

(1) The number of days of the bleeding episode with at least 'mild' bleeding will be derived.

(2) The total MBL will be divided by the number derived in (1).

(3) The 'mean' MBL derived in (2) will be added to the AH values of all days of the bleeding episode with at least 'mild' bleeding. This step will be performed after having imputed missing bleeding intensities but before having imputed missing AH values (cf. Section 4.3.1). The resulting AH values will not be used for the imputation strategy of missing AH values described in Section 4.3.1.

4.3.3 Ultrasound-related volume measurements

Partially missing data in ultrasound measurements for determining fibroid volume, uterine volume and cyst volume will be handled as described in Section 4.5.

4.3.4 Last study drug intake on eCRF Study Drug Exposure page

Missing dates of last study drug intake on eCRF study drug exposure page used for defining

- the end of a treatment period and the reference end date of treatment phase will be imputed by the date of last study drug intake in the same treatment period reported in the eDiary;
- treatment duration and compliance to study drug will not be imputed (cf. Section 6.1.6).

4.4 Interim Analyses and Data Monitoring

With the implementation of the Integrated Clinical Study Protocol, version 6.0 a safety and efficacy analysis is planned after all subjects have completed their treatment period. All data until 31 DEC 2019 will be cleaned and included in the analysis. The complete efficacy data used for the confirmatory analysis has been collected until this point in time and will be used for the efficacy analysis. Therefore, the confirmatory analysis will be final at this point in time. Additionally, available data up to the time point of analysis will be included in this analysis.

Following this analysis neither the sample size nor the study design will be adjusted. The reason for this is that the confirmatory efficacy analysis will only be performed once and none of the analysis results will trigger decisions on the further conduct of the study, such as an early termination. Therefore, no alpha adjustment or power consideration are required. In addition, no Data Safety Monitoring Board (DSMB) will be required.

The final analysis will follow the end of the study safety closeout visit.

4.5 Data Rules

This section outlines definitions of relevant variables as well as rules for deriving variables. Variables are sorted alphabetically.

Baseline cycle: First day of the baseline cycle is day one of the bleeding episode following Visit 1. The last day of the baseline cycle is the day preceding the next bleeding episode. The baseline cycle length is defined as the number of days from the first day of cycle until the last day of cycle. These definitions will apply if during this cycle the inclusion criterion of HMB diagnosis is fulfilled (MBL >80.00 mL). If HMB is not confirmed in the first bleeding episode following Visit 1 then the second bleeding episode recorded during screening will be considered. Similarly, the first day of the second cycle is the day one of the second bleeding episode following Visit 1 and the day preceding the next bleeding episode will be considered the last day of the cycle. Again, this second bleeding episode will be considered the baseline cycle only if the HMB criterion is fulfilled, otherwise the next bleeding episode, if recorded, will be inspected for the diagnosis of HMB. If none of the bleeding episodes preceding the first drug administration fulfills the HMB inclusion criterion the first bleeding episode following Visit 1 will be considered for the baseline cycle. If no such bleeding episode exists, then the bleeding episode already ongoing at Visit 1, if recorded, will be considered for the baseline cycle, starting with Day 1.

Note that during the screening period the diagnosis of HMB will be inspected using the definition of bleeding episode based on the AH method, as explained in this section and in the protocol. In contrast, during the analysis only the definition of bleeding episode based on the UF-DBD will be used. Therefore, the baseline cycle defined during the analysis here described is independently derived and may differ from the bleeding episode used to assess the HMB diagnosis during the screening period.

Baseline menstrual blood loss (MBL): The baseline MBL is defined as the sum of MBL volume from all days included in the baseline cycle. If baseline cycle length is more than 28

days, only data from the first day of cycle to the 28th day of cycle will be used. If baseline cycle length is less than 28 days or equal to 28 days, data from the first day of cycle to the last day of cycle will be used.

Baseline values and changes from baseline: If an analysis is performed by visits or by Intervals 'Exit Examination' (see *Time Intervals*), the last valid, non-missing value prior to the Reference start date will be considered as baseline value. If the analysis is performed by any time interval different than Intervals 'Exit Examination', the 'worst' measurement will be used, whereby 'worst' is defined for the considered parameter within the specific section.

(*Absolute) changes from baseline* will be calculated as the difference between the post-baseline value and the baseline value:

Absolute change = post baseline value – baseline value.

Relative changes from baseline, also referred to as *percent change from baseline*, are derived for some variables. It is defined as

Relative change = 100 × [post baseline value – baseline value] / baseline value.

Those baseline definitions apply to the analysis of all TPs. For variables analyzed by 28-day periods, the baseline period corresponds to the last 28 days before Reference start date. This definition does not apply to baseline MBL volume where a specific definition is provided in this section (see *Baseline menstrual blood loss)*.

Bleeding episode (assessed by the UF-DBD): A bleeding episode is defined as day(s) with bleeding/'spotting' of which at least one day is of intensity 'mild' or higher, preceded and followed by at least 2 consecutive days with eDiary entry 'no vaginal bleeding'. The preceding 2 bleeding-free days may not be recorded for the first bleeding episode at study entry [1].

This definition of bleeding episode, based on the UF-DBD, will be implemented in the analysis to identify bleeding episodes. The protocol includes a definition of bleeding episode based on the AH method, which will be implemented during the screening period to facilitate the assessment of the HMB inclusion criterion (criterion 5, MBL >80 mL) by clinical sites. This latter definition of bleeding period will not be used in the statistical analysis.

Break period: The break period is defined as the treatment-free period between TP1 and TP2. The break period starts the day following the date of stop date of study drug administration in TP1, and ends the day before the study drug administration starts in TP2. Prior to the Global Protocol Amendment, version 4.0, subjects were instructed to start TP2 within Days 3 to 7 of the first bleeding episode following the end of TP1. If a subject did not start TP2 (due to the temporary pause or for any other reason), the subject does not have any break period but only a follow up period which starts at the first day after the end of TP1. For efficacy variables evaluated within the break period, part of the data from the follow-up period of those patients will be included in the analysis of the break as well. In this case a theoretical end date which reflects the assumed end date of the break period if the subject had started TP2 will be generated. For this purpose information on the start of TP1 will be used: If a subject started TP1 at the i-th Day of her bleeding episode, the theoretical start date of TP2 will be assumed as the i-th Day of her first bleeding episode during treatment break. The theoretical end date of the break period will then be assumed as the day before the theoretical start date of TP2.

Endometrial biopsy conducted: A biopsy will be regarded as conducted if the intervention was attempted, either successfully or unsuccessfully. Successful means that the date of biopsy is given and 'Biopsy sample obtained - Not done' is not ticked in the eCRF. Unsuccessful means that the date of biopsy is given, 'Biopsy sample obtained - Not done' is ticked in the eCRF, and the reason 'Unsuccessful attempt' is selected.

Largest fibroid: The largest fibroid is defined as the fibroid with the largest volume at baseline (measured by ultrasound). The fibroid identified at baseline as the largest will be followed up at subsequent visits.

Menstrual blood loss (MBL) for 28-day periods: The MBL for a 28-day period is defined as the sum of MBL (in mL) assessed by AH/MP method in a 28-day period. Non-overlapping, consecutive 28-day periods will be created along the different study periods: screening, treatment periods, break period and FUP. For details about the volume of MBL assigned to MP, see Appendix 1.

Reference start/end dates of treatment phase: The reference start and end dates of treatment phase refer to the first and last study drug intake and will be identified based on the eCRF entry (EXCATN=1). Missing eCRF end dates will be replaced by the date of the last study drug intake according to the eDiary. In case there is no study drug intake recorded in the eDiary, the date of the last available visit will be used, which is the maximum of the last date at which a visit took place (visit end date), and the "last visit" as collected on the eCRF page.

Repeated measurements at the same visit: If more than one post-randomization measurement is available for a given visit, the first observation will be used in the analysis if no special reason for the additional observation was provided in the eCRF.

All observations will be presented in data listings.

Time intervals: Due to the complex visit schedule of the studies, a time-interval based approach is chosen to present data over time. Several time intervals are defined which are used for different analysis topics ('BL/Post baseline', 'Intervals A', 'Intervals B', 'Intervals C', 'Intervals D', 'AE Intervals', 'Exit Examination'). Those are depicted in Figure 4–1. The selection of time intervals used for a certain analysis topic will be described within each section.

Figure 4–1: Overview of time intervals used for the statistical analysis.

BL/Post baseline	BL			Post baseline															
Intervals A	BL			Treatment Phase									FUP Phase						
Intervals B	BL			TP1 + BR						TP2			FUP1+2						FUP ...
Intervals C	BL			TP1			BR			TP2			FUP1			FUP2			FUP ...
Intervals D	BL 1-28	BL 29-56	BL 57-x	TP1 1-28	TP1 29-56	TP1 57-84	BR 1-28	BR 29-56	BR ...	TP2 1-28	TP2 29-56	TP2 57-84	FUP 1-28	FUP 29-56	FUP 57-84	FUP ...	FUP ...	FUP ...	FUP ...	FUP ...
AE Intervals	Pre-treatment			"on" (TP1 + 8d)			"off"			"on": TP2 + 8d			"off"			Post-treatment (Start: day 61 after EoT)			
Exit Examination	BL, last value within interval												Exit examination (Start: EoT -7 days); last value within this interval						

Abbreviations: BL: Baseline, TP: Treatment Period, BR: Break, FUP: Follow-up, EoT: End of treatment, d: days.

Treatment periods and breaks will be identified as described above. Baseline/Pre-treatment starts with the date of informed consent and ends on the day before the start of the treatment phase. The treatment phase starts and ends with the reference start and end date. The FUP phase starts on the day after the end of the treatment phase and ends on the maximum of the last date at which a visit took place (visit end date), and the "last visit" as collected on the eCRF page. Baseline, treatment phase, FUP phase, treatment periods and breaks are subdivided into subsequent periods of different lengths. "Exit examination" intervals include the baseline measurement, i.e. the last non-missing value before reference start date, and the exit examination which is defined as the last non-missing value at or after the reference end date minus 7 days. It is per definition the safety closeout visit for subjects who perform this visit.

Treatment period: Each treatment period is planned to consist of three subsequent 28-day periods. In case of premature discontinuation of study drug the subject's respective treatment period might be shorter than 84 days. If a subject took the study drug for more than 84 days of a treatment period, the respective treatment period will be longer than 84 days, in contrast. Treatment periods will be numbered sequentially.

Start and end of a treatment period will be identified based on eCRF entries (EXCATN=1). The start of TP1 will be the reference start date (i.e. the day of the first study drug intake). The end of TP1 will be the date of the last study drug intake for this treatment period. The start of TP2 will be the date of the first study drug intake in TP2. The end of TP2 will be the date of the last study drug intake in TP2. This coincides with the reference end date if the subject has started TP2.

If the last study drug intake within TP1 is missing in the eCRF, then the date of the last study drug intake within the respective treatment period according to the eDiary (EXCATN=5 and EXMEDN≥1) will be used for defining the end date of the treatment period.

Replacement rules for missing end date of the last treatment period are analogue to those described for reference end dates.

Ultrasound-related measurements: The derivation of fibroid volume, uterine volume and cyst volume based on ultrasound measurements is described in the following.

- *Fibroid volume*

  A fibroid's volume will be calculated by $\frac{\pi \times a \times b \times c}{6}$ (cf. [2]), where

  a = largest diameter of the fibroid (cm),

  b = largest diameter perpendicular to the largest diameter, a, (cm),

  c = largest diameter perpendicular to first and second largest diameters, a and b, (cm),

  with a, b and c measured by ultrasound.

  If only two diameters are available, the missing diameter will be replaced by the mean of the available two diameters and the fibroid's volume will be calculated as described above.

  If only one diameter is available, the fibroid's volume will be calculated by $\frac{4 \times \pi \times r^3}{3}$, where

$r = \frac{available\ diameter\ of\ the\ fibroid\ (cm)}{2}$.

- *Uterine volume*

  The uterine volume (mL) will be calculated by $\frac{\pi \times a \times b \times c}{6}$, where

  a = maximum width anteroposterior of uterus (cm),

  b = maximum width transverse of uterus (cm),

  c = corpus length + cervix length (cm),

  with *a*, *b* and *c* measured by ultrasound.

  If only two or one of *a*, *b*, and *c* is available, the uterine volume will be derived based on only the available measurements as has been described for the fibroid volume.

- *Cyst volume*

  The volume of a cyst will be calculated by $\frac{\pi \times a \times b \times c}{6}$, where

  a = largest diameter of the cyst (cm),
  b = largest diameter perpendicular to the largest diameter of the cyst (cm),
  c = $\frac{(a+b)}{2}$ (cm),

  with *a* and *b* measured by ultrasound.

  If only one diameter is available, the volume of the cyst will be calculated by $\frac{4 \times \pi \times r^3}{3}$, where

  $r = \frac{available\ diameter\ of\ the\ cyst\ (cm)}{2}$.

4.6 Blind Review

Important deviations from the protocol and validity findings and the resulting assignment of subjects to the analysis sets (see Section 5.1) are agreed upon in the blind review meeting (BRM). The documentation of important deviations, validity findings and the assignment of subject data to analysis sets will be performed according to the sponsor's applicable Standard Operating Procedures and/or Instruction Manuals. The definition for important deviations and validity findings will be provided in the 'Specification of assessment criteria and identification requirements' before unblinding the data.

Identification of important deviations and validity findings will be done periodically while the study is running, concluding with the completion of the final list during the BRM. Any changes to the statistical analysis prompted by the results of BRM will be documented in an amendment and, if applicable, in a supplement to this SAP.

The list of important deviations and validity findings will only be final after assessment of conditional findings, which are validity findings that can only be identified after unblinding of study treatment.

5. Analysis Sets

5.1 Assignment of analysis sets

Final decisions regarding the assignment of subjects to analysis sets will be made during the blind review of study data and documented in the final list of important deviations, validity findings and assignment to analysis set(s) (see Section 4.6).

The following statistical analysis sets will be defined:

Full analysis set (FAS): All randomized subjects, excluding randomized subjects who did not start TP1 due to the study being temporarily paused. Subjects will be analyzed as randomized.

Per protocol set (PPS): All subjects in the FAS without any validity findings impacting the primary efficacy variable in TP1.

Safety analysis set (SAF): All subjects who took at least 1 dose of study drug based on 'Study Drug Exposure' eCRF page. Subjects will be analyzed as treated. Mistakes in the administration of study drug may affect the assignment to treatment. In general, subjects who receive a different study drug to the one they were randomized to will be assigned to the treatment most often received.

The primary efficacy variable will be analyzed on the FAS and PPS, whereby the analysis on the FAS is considered to be the primary one. Safety analyses, medical history and concomitant medication will be performed on the SAF. Demographic data, baseline characteristics, exposure and compliance will be analyzed on the FAS, SAF and PPS. The FAS will be used for the display of all other variables. Analyses on the PPS will be additionally conducted as sensitivity analyses for selected secondary efficacy variables.

All screened patients who are not included in the FAS, will be used solely in disposition tables, and will be listed only in Section 16 of the clinical study report, as described elsewhere.

6. Statistical Methodology

6.1 Population characteristics

In general, descriptive statistics will be presented for variables defined in this section. For continuous variables, number of observations, mean, standard deviation, minimum, median, and maximum will be presented. For categorical variables, number and percentage of subjects will be presented. Listings will be provided as appropriate.

6.1.1 Subject validity and disposition

The following data related to subject validity and disposition will be summarized overall and/or by treatment group, if applicable and not specified otherwise:

- Screening failures
- Number of subjects enrolled – overall, by country and by study site
- Number and percentage of subjects randomized,

  - who received at least one dose of study drug (based on 'Study Drug Exposure' eCRF page),
  - for whom study drug was never administered (based on 'Study Drug Exposure' eCRF page),
  - who completed and did not complete the study, whereby completed refers to the completion of all phases of the study including the 'last visit' or 'the last scheduled procedure shown in the Schedule of Activities'.
- Number and percentage of subjects who completed and discontinued each epoch with the reasons for discontinuation
- Number and percentage of subjects who discontinued study drug with the reasons for discontinuation
- Subject validity and primary reasons for exclusion from analysis set
- Number and percentage of subjects who performed adrenal monitoring, skin monitoring and endometrial monitoring within the safety closeout visit, and further details about the participation in the safety closeout visit.

A listing of the subjects' assignment to the FAS, SAF and PPS and the reasons for exclusion will also be provided by treatment group.

6.1.2 Demographics and baseline characteristics

All demographic and baseline characteristics will be summarized by treatment group (including pooled treatment group A1+B2). The descriptive statistics will be presented for the FAS, SAF and PPS.

Demographic and baseline assessments to be summarized will include:

- Age (at inclusion), region/country, race, ethnicity
- Weight (kg), height (cm), body mass index (BMI; kg/m^2)
- Categorized BMI (< 18.5, 18.5 to < 25, 25 to < 30, ≥ 30 kg/m^2)
- Smoking and alcohol consumption
- Educational level (i.e., level of education, years of school and professional education)
- Baseline MBL (by AH method/MP) and categorized MBL group ($\leq$ 150 mL, >150 to 300 mL, > 300 to 500 mL and > 500 mL)
- Volume of 3 largest fibroids and location of largest fibroid (International Federation of Gynecology and Obstetrics [FIGO] classification) by ultrasound (mL)
- Largest diameter of fibroid by ultrasound (mm)
- Uterine volume by ultrasound (mL)
- Endometrial thickness by ultrasound (mm)
- Hemoglobin (g/dL)

Demographic and baseline characteristics will be summarized also for the following subgroups:

- By region/country (Japan, US, other countries)
- By ethnicity
- By race

6.1.3 Medical history

For medical history the Medical Dictionary for Regulatory Activities (MedDRA; current version at the time of analysis) will be used. Medical history findings (i.e., previous diagnoses, diseases or surgeries) not pertaining to the study indication, starting before signing of the informed consent and considered relevant to the study will be tabulated by primary system organ class (SOC) and preferred term (PT) by treatment group. Medical history will be presented for SAF.

All new or worsened findings after signing the informed consent should be documented on the AE eCRF page.

6.1.4 Reproductive, menstrual, and fibroid history

Reproductive and menstrual history will include information on age at menarche (in years), number of pregnancies, number of births, years since last birth or abortion and inability to conceive (at any time and during the last 1 year) and duration of unsuccessful attempts (in months).

Fibroids history will include information on family history, time since initial diagnosis (in months), and reason for initial diagnosis, current and past symptoms, and previous medical treatments and procedures relevant for uterine fibroids, if applicable.

In addition, any performed procedures for uterine fibroids or symptoms of uterine fibroids and the consideration of surgical procedures within the next 12 months in case of remaining symptoms (as they are at screening) will be summarized.

These variables will be analyzed descriptively, separately for each treatment group based on the SAF.

6.1.5 Heavy menstrual bleeding questions

This set of questions has been developed as a tool to identify women with HMB. It will be used at Visit 1 and the responses can be entered directly into electronic data capturing system RAVE, which will be considered as primary source data. The questionnaire includes 14 questions in total. The response (yes/no) to each question will be presented descriptively in frequency tables separately for each treatment group based on the SAF.

6.1.6 Exposure and compliance to study drug

Exposure, treatment duration and compliance to study drug will be analyzed descriptively based on the SAF, FAS and PPS. They will be presented by treatment period and overall, (i) by treatment group and (ii) by study drug. Overall treatment duration will be shown including and excluding treatment break. Both study drug exposure and compliance will be calculated

separately based on the eDiary data as well as on the eCRF (namely the 'Study Drug Exposure' and the 'Drug Accountability' pages).

Subject's *exposure* to the study drug will be calculated as total number of tablets taken during a time frame (i.e. either a specific treatment period or both treatment periods combined), using information from the 'Drug Accountability' eCRF page and the subject's eDiary. Missing numbers of tablets taken based on eCRF page will not be replaced, while missing numbers of tablets taken based on eDiary will be replaced with 0.

The *treatment duration* is defined as the number of days from the day of first study drug intake up to and including the day of last study drug intake in the treatment period of interest. Treatment duration will be calculated separately for the two treatment periods. In addition, it will be computed for both treatment periods combined excluding the time for treatment break and for the treatment phase (i.e. including the break). For the latter, treatment duration per specific study drug is calculated from start of TP1 until start of TP2 and from start of TP2 until end of TP2 (i.e. the break is included in the duration of TP1). If a patient discontinued before TP2, treatment duration is calculated from start to end of TP1.

The days of first and last study drug intake will be taken from the 'Study Drug Exposure' eCRF page. Missing dates in the eCRF will not be replaced.

Treatment compliance of a subject is defined as the total number of tablets taken during a specific time frame (i.e. the exposure) divided by the treatment duration (in days) for that time frame, and will be given in percent. Compliance calculation is derived from exposure based on (i) the eCRF and (ii) eDiary and treatment duration (excluding breaks) based on the eCRF, as described above.

Compliance will be derived as follows:

$$Compliance = \frac{Number\ of\ tablets\ taken}{Treatment\ duration} \times 100\%$$

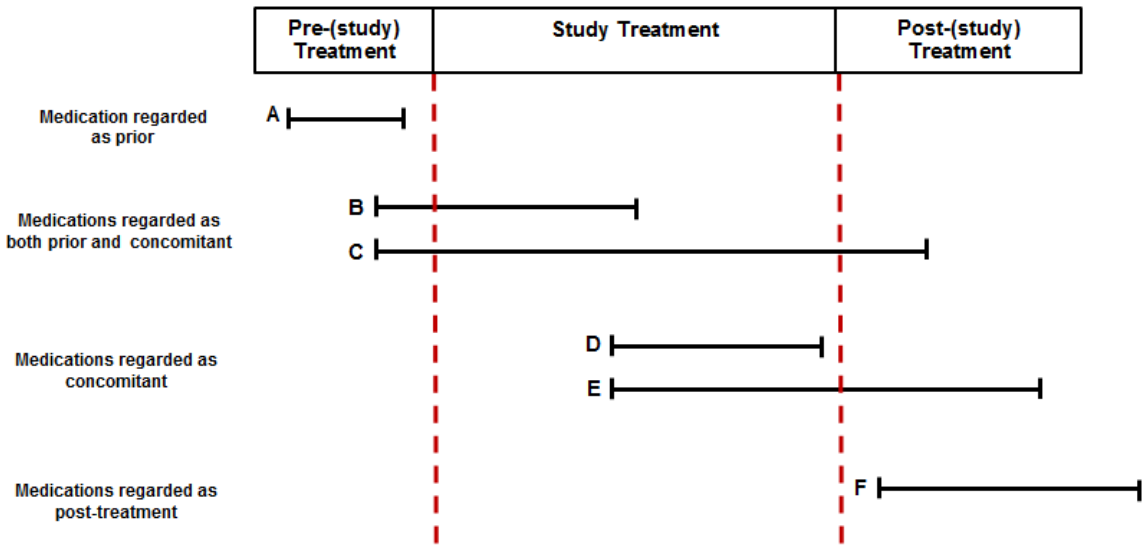
6.1.7 Prior, concomitant and post-treatment medication

For prior, concomitant and post-treatment medications, the following definitions in accordance with the Global Standards Catalogue V4.0 will be used in the analysis (see Figure 6–1):

- *Prior medication*: Medication taken before start of the study drug intake, i.e. before start of TP1 (regardless of when it ended).
- *Concomitant medication*: Medication taken during treatment phase, i.e. between first and last study drug intake (regardless of when it ended).
- *Post-treatment medication*: Start of medication is after the treatment phase, i.e. after last study drug intake.

In case it's not clear if a medication is prior, concomitant or post-treatment (for example, due to partially missing dates), medication will be classified as "concomitant".

Figure 6–1: Categories of medication (example)



Categories are prior medication (A, B, C), concomitant medication (B, C, D, E) and post-treatment medication (F). Source: Global Standards Catalogue V4.0

Medication, recorded as prior, concomitant or post-treatment medication in the eCRF, will be coded according to the World Health Organization Drug Dictionary WHODRUG GLOBAL B3 (initially September 1, 2017, within Bayer referred to as '2017SEP'), to the respective Drug Codes with their corresponding Anatomical Therapeutic Chemical (ATC) classification.

The number of subjects taking prior, concomitant or post-treatment medication will be analyzed using frequency tables. Analysis of prior, concomitant and post-treatment medication will be done on the SAF. Prior, concomitant and post-treatment medication will be shown by treatment group.

6.2 Efficacy

The following sections describe the efficacy analyses of primary, secondary, and other efficacy variables in detail. Further variables that were captured during the study will be either analyzed descriptively or displayed in subject listings.

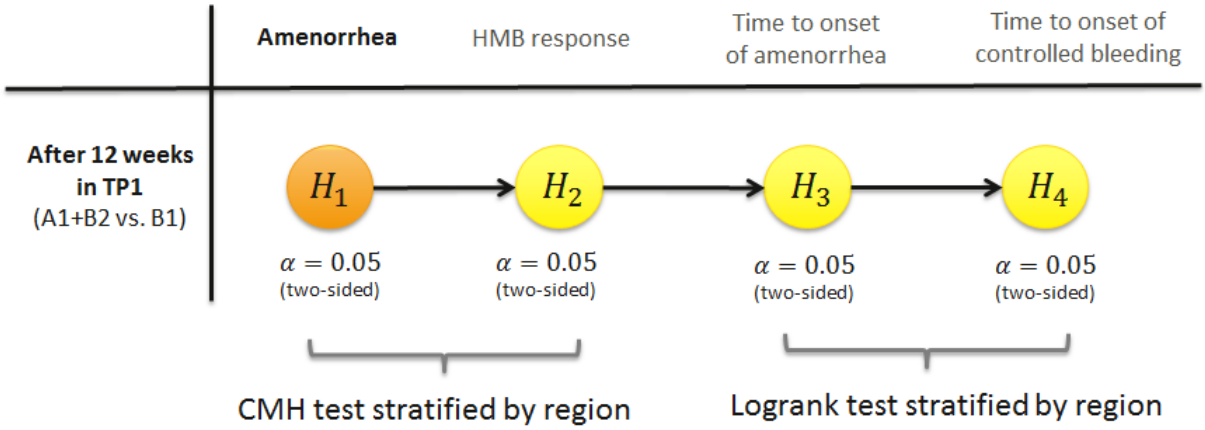
The primary efficacy variable and secondary efficacy variables will be calculated based on the AH method. Other efficacy variables related to MBL will be calculated based on both the AH method and MP. If an endometrial biopsy was conducted during a time period for which bleeding will be assessed, bleeding or 'spotting' on the day of biopsy and the 3 days thereafter will not be considered in the derivation of any variables that belong to primary, secondary or other efficacy variables. Thus, prior to the assessment of bleeding variables, such as amenorrhea, for example, the AH and UF-DBD entries during those days will be set to 0 mL and "No vaginal bleeding", respectively.

For the confirmatory efficacy analysis, a hierarchical testing approach will be applied, involving the primary efficacy variable amenorrhea (yes/no) and the first three secondary efficacy variables HMB response, time to onset of amenorrhea, and time to onset of controlled bleeding. These tests always include a comparison of vilaprisan 2 mg versus placebo:

> After 12 weeks of treatment in TP1:
> Comparison of vilaprisan 2 mg in pooled treatment groups A1 and B2 (as the treatments are the same and pooling increases the power) vs. placebo in treatment group B1.

In total, four tests will be carried out each to an alpha level of 0.05, see also Figure 6–2: First, the test for the primary efficacy variable amenorrhea (yes/no) will be carried, followed by the test for HMB response (yes/no), the test for time to onset of amenorrhea, and finally, the test for onset of controlled bleeding. As the hierarchical testing procedure follows a fixed sequence, it stops as soon as any of these tests cannot be rejected to an alpha level of 0.05, and all further tests after failing to reject one null hypothesis in the testing sequence will be considered exploratory. This fixed sequence procedure accounts for the multiplicity created by carrying out multiple tests.

Figure 6–2: Hierarchical testing strategy for the confirmatory efficacy analysis



Abbreviations: HMB: Heavy menstrual bleeding, CMH: Cochran-Mantel-Haenszel, H: Hypothesis to be tested related to the amenorrhea (H_1), HMB response (H_2), time to onset of amenorrhea (H_3) or time to onset of controlled bleeding (H_4), TP: Treatment period, A1: Vilaprisan 2mg; 2 treatment periods of 12 weeks, B1: Placebo; 1 treatment period of 12 weeks and Vilaprisan 2mg; 1 treatment period, B2: Vilaprisan 2mg; 1 treatment period of 12 weeks and placebo; 1 treatment period.

The study will be considered successful if at least superiority of vilaprisan 2 mg vs. placebo in Treatment Period 1 (TP1) based on the primary efficacy variable can be demonstrated. In general, superiority of vilaprisan treatment is shown if the respective null hypothesis is rejected to a local alpha level of 0.05 and point estimates of treatment effect favor vilaprisan.

The primary analysis will be conducted based on the FAS (see Section 5.1).

6.2.1 Primary efficacy variable

The primary efficacy variable is amenorrhea (yes/no), defined as MBL <2 mL during the last 28 days of treatment. Evaluation of MBL will be based on AH method for a treatment period of interest.

The primary efficacy variable will be derived after the imputation of missing values described in Section 4.3.

If after the imputations described in Section 4.3 and considering potential biopsies, neither of the AH values for the last 28 days are missing, AH values and bleeding intensities of the last 28 days of a treatment period are used to derive the primary efficacy variable.

If, in contrast, there are AH values in the last 28 days of treatment which are still missing (i.e., where an imputation was not possible), it is further studied whether the corresponding bleeding intensities of the UF-DBD are missing on the same day(s) as well.

(a) If the bleeding intensity is not missing and only the AH value is missing, the bleeding intensity must be at least 'mild', because for 'no vaginal bleeding' and 'spotting' the AH value would have been imputed with 0 mL. Accordingly, it must be assumed that bleeding has occurred on these day(s). Therefore, the subject is considered as not having amenorrhea. Step b) will then not be applied irrespective of whether there are also days for which both, the bleeding intensity and the AH value, are missing.

(b) If both, the bleeding intensity and the AH value, are missing for any of the last 28 days, the 28-day period will be extended: The left margin of the 28-day period will be displaced backwards in time and step-wise until the required number of 28 non-missing AH values are included. The left margin may be displaced backwards at maximum up to the end of the first bleeding episode when treatment was started in respective treatment period.

Considering this time window, a subject will be considered as not having amenorrhea if

i) the sum of the non-missing (i.e. original or imputed) AH values is ≥ 2 mL, or

ii) there is at least one day with missing AH value and bleeding intensity of 'mild' up to 'very severe' after having applied the imputation rules, or

iii) the subject did not complete at least 8 weeks of treatment.

If neither of i), ii) or iii) applies, the subject will be considered as having amenorrhea.

6.2.1.1 Confirmatory analysis of the primary efficacy variable

The primary efficacy variable will be analyzed descriptively in frequency tables by treatment period and region within each treatment group. Additionally, vilaprisan treatment after 12 weeks, i.e. in TP1, of treatment groups A1 and B2 will be pooled.

The confirmatory analysis will be performed only based on TP1. In the confirmatory analysis, the proportion of subjects with amenorrhea will serve as comparison criteria in the comparisons defined in Section 6.2. For simplicity, the confirmatory analysis will be formulated generally. The null hypothesis, H_0, states that the common difference of proportions between the 2 mg vilaprisan group and the placebo group among the regions (US, Japan and other) δ is zero. Considering a two-sided test problem with hypotheses

$$H_0: \delta = 0 \qquad \text{vs.} \qquad H_1: \delta \neq 0,$$

the alternative hypothesis, H_1, states that the common difference of the proportion of subjects with amenorrhea between the 2 mg vilaprisan group and the placebo group among the regions is unequal to zero.

The test problem will be investigated by means of a two-sided, continuity-corrected Cochran-Mantel-Haenszel (CMH) test stratified by country/region at a local 0.05 significance level. The test statistic of the CMH test is defined as

$$CMH = \frac{\left[\left|\sum_k \left(n_{11k} - \mathrm{E}_{H_0}(n_{11k})\right)\right| - 0.5\right]^2}{\sum_k \mathrm{Var}_{H_0}(n_{11k})},$$

with n_{ijk} representing the number of subjects with ($i = 1$) or without amenorrhea ($i = 0$) and belonging to the vilaprisan group ($j = 1$) or to the placebo group ($j = 0$) and country/region k (1 ≙ US, 2 ≙ Japan, 3 ≙ other countries). It compares the observed number of subjects with amenorrhea in the vilaprisan 2 mg group with the expected number under the null hypothesis [3]. Under the assumption of fixed $n_{1+k} = \sum_{j=0}^{1} n_{1jk}$, the expected number and the variance under the null hypothesis are [4]

$$\mathrm{E}_{H_0}(n_{11k}) = n_{1+k} n_{+1k} / n_{++k}$$

and

$$\mathrm{Var}_{H_0}(n_{11k}) = \frac{n_{1+k} n_{0+k} n_{+1k} n_{+0k}}{{n_{++k}}^2 (n_{++k} - 1)}$$

The plus sign (+) denotes a summation over the respective index.

Under the said assumption and the null hypothesis, the CMH test-statistic approximatively follows the χ^2-distribution with one degree of freedom [3]. Hence, the null hypothesis of conditional independence between the treatment and the presence of amenorrhea given the country/region will be rejected, if the observed test-statistic exceeds the 95% quantile of the χ^2-distribution with one degree of freedom, which is 3.841.

Along with the results of the CMH test, the Mantel-Haenszel estimate for the common risk difference of the proportion of subjects with amenorrhea between the 2 mg vilaprisan group and the placebo group among the regions δ will be reported together with its approximate 95% confidence interval (two-sided).

The Mantel-Haenszel estimator for the common risk difference

$$\hat{\delta} = \frac{\sum_k \hat{\delta}_k w_k}{\sum_k w_k},$$

is composed of the estimated risk difference of the proportion of subjects with amenorrhea in country/region *k*,

$$\hat{\delta}_k = \frac{n_{11k}}{n_{+1k}} - \frac{n_{10k}}{n_{+0k}},$$

and country/region-specific weights

$$w_k = \frac{n_{+1k}\, n_{+0k}}{n_{++k}}$$ [5].

An approximate $100(1-\alpha)\%$ confidence interval (two-sided) for the common risk difference $\hat{\delta}$ is given by

$$\left[\hat{\delta} \pm z_{\alpha/2}\sqrt{\widehat{Var}(\hat{\delta})}\right],$$

where $z_{\alpha/2}$ denotes the $(\alpha/2)$-quantile of the standard normal distribution and the variance estimator of the Mantel-Haenszel estimator $\hat{\delta}$ will be computed as

$$\widehat{Var}(\hat{\delta}) = \frac{\hat{\delta}\sum_k u_k + \sum_k v_k}{(\sum_k w_k)^2},$$

where

$$u_k = \left[n_{+1k}^2 n_{10k} - n_{+0k}^2 n_{11k} + \frac{n_{+1k}n_{+0k}(n_{+0k} - n_{+1k})}{2}\right] / n_{++k}^2$$

and

$$v_k = \frac{n_{11k}(n_{+0k} - n_{10k}) + n_{10k}(n_{+1k} - n_{11k})}{2n_{++k}}$$ [6].

6.2.1.2 Sensitivity analyses of the primary efficacy variable

6.2.1.2.1 Analysis in PPS population

As a first sensitivity analysis, the impact of the choice of analysis set will be investigated. The confirmatory analysis of the primary efficacy variable will thus be repeated on the PPS population excluding subjects with validity findings impacting the assessment (see Section 5.1) of amenorrhea during the last 28 days of TP1. As opposed to the analysis on the FAS, this is intended to measure the actual effect in protocol adherers rather than investigating the effect of the treatment strategy in every randomized subject.

6.2.1.2.2 Analysis based on alternative missing data imputation strategy

The impact of missing data will be investigated in the FAS (according to the intention to treat principle) with multiple imputation, using a pattern-mixture framework as described by the National Research Council (NRC 2012) [7]. In the primary analysis, it is assumed that values in the UF-DBD or for the AH method are missing at random (MAR), i.e., they only depend on observed outcomes and can thus be imputed based on the observed data. However, it is possible that missing values are systematically different from observed values because patients are dissatisfied with their treatment and are consequently more likely to withdraw from the study, for example. If the differences between missing and observed values do not balance out between the treatment groups, then this creates a bias in the treatment effect. Therefore, this sensitivity analysis serves to examine the missing data mechanism and explore different scenarios under a MNAR (Missing Not at Random) assumption where an outcome depends on the fact whether it is observed or not and missing outcomes can thus differ systematically from observed outcomes by treatment group.

For the purpose of this sensitivity analysis for missing data, the standard procedure of imputing intermittent single or two consecutive days with missing UF-DBD bleeding intensities and the replacement of missing daily AH values with 0 mL for bleeding intensities of not more than 'spotting' are kept as sensible approaches because nothing out of line is expected to happen in the time frame of one or two days. In contrast to the primary analysis, however, days with biopsies and the three days thereafter will not be considered with an AH value of 0 mL (assuming no bleeding), but will be considered missing since the AH values on those days if the patient had not had the biopsy are unknown. This rule and further rules used to impute missing data in the primary analysis will be challenged with this analysis using multiple imputation in a pattern-mixture approach.

Subjects are sorted into three categories based on the available UF-DBD and AH information from Day 57 to 84 of the TP1 (or the last 28 days of TP1, if study drug was taken for more than 84 days). Those three categories are comprised of clear responders and clear non-responders (after the initial minimal imputation rules described above) as well as subjects with unclear response status. A clear responder for amenorrhea is defined as a subject

- who has only UF-DBD bleeding intensities of no more than 'spotting' AND
- for whom the sum of daily AH values is less than 2 mL AND
- for whom there are no missing values for either the AH method (daily values) or the UF-DBD bleeding intensities.

As opposed to this definition, a clear non-responder for amenorrhea is defined as a subject

- who has at least one bleeding intensity of 'mild' or higher OR
- for whom the sum of daily AH values is equal to or higher than 2 mL.

All response criteria are set in relation to the planned time frame for the assessment of the primary efficacy variable, i.e., Day 57 to Day 84 of the TP1. An exception is made when a subject's TP1 is longer than 84 days. In this case the last 28 days of TP1 are used for the assessment.

A subject is considered as unclear responder for amenorrhea (and thus having a missing outcome) if in the relevant time frame, all the subject's bleeding intensities do not exceed 'spotting' and the sum of daily bleeding intensities is below 2 mL, but there are still missing values among both the UF-DBD bleeding intensities and the daily AH values between Day 57 and 84 of the TP1. Therefore, it is still possible that replacement of the missing bleeding information turns an apparent responder into a non-responder for amenorrhea, signifying that no clear assessment about the response status can be made.

A descriptive analysis of the number of subjects in each response category will be supplied by treatment group.

Missing data imputation will only be applied to subjects with unclear response status, whereas not the AH values will be imputed but the response status. Response status of subjects with unclear outcome are generated in a three-step process [8] by

1) Estimation of individual log-odds of amenorrhea for all patients from a logistic regression model, which is an imputation model fitted based on data of clear responders and non-responders,
2) Simulation of response status using the predicted log-odds to create multiple datasets with imputed data; conduct of the Cochran-Mantel-Haenszel test on each imputed dataset; combination of the analysis results to generate statistical inference,

3) Assessment of the robustness of step 2 after adding/subtracting a penalty (shift parameter) from the log-odds in the placebo group and vilaprisan group, respectively, and determination of the "tipping point".

Those three steps are outlined in more detail in the following.

Step 1:

To impute the missing response status, a logistic regression model will be fitted to the binary variable amenorrhea yes/no (coded as 1 and 0) for subjects who are either clear responders or clear non-responders. Treatment group (X_T), stratification factor region (X_R), baseline covariates age (X_{Age}) and baseline MBL (X_{BMBL}) are used as covariates. This leads to the following model equations for the log-odds $logit(p_1) = \log(p_1/(1-p_1))$ of the response (amenorrhea=yes) probability p_1:

$$logit(p_1) = \beta_0 + \beta_1 X_T + \beta_2 X_R + \beta_3 X_{Age} + \beta_4 X_{BMBL}.$$

If the model does not converge, the stratification factor region will not be used as covariate in the model.

Step 2:

Following the multiple imputation approach with the monotone logistic regression method [9], new parameters $\beta_* = (\beta_{*0}, \dots, \beta_{*4})'$ are drawn from the posterior predictive distribution of the regression parameters $\beta = (\beta_0, \dots, \beta_4)'$. For a subject with unclear response status and covariate vector $x = (1, x_T, x_R, x_{Age}, x_{BMBL})'$ the predicted response probability of amenorrhea is given by

$$p_1 = \frac{\exp(x\prime\beta_* + \delta)}{1+\exp(x\prime\beta_* + \delta)},$$

where δ denotes the penalty or shift parameter which is 0 under the MAR assumption. A random variable u is drawn from a standard uniform distribution and if the unequality $u < p_1$ holds for a subject then the subject's response status for amenorrhea is imputed with yes (=1), else it is imputed with no (=0). After imputing all subjects' response statuses in this manner, the primary analysis is repeated, i.e., the Cochran-Mantel-Haenszel test is conducted based on the complete imputed data set. This imputation and analysis will be repeated 1,000 times. Inferences for combined parameters will be done using Rubin's multiple imputation rules [10] which reflect imputation uncertainty.

Step 3:

To assess the robustness of the analyses for deviations from the MAR assumption, step 2 will be repeated with different choices of the shift parameter for subjects from the vilaprisan group and placebo group, respectively. To explore the whole range of possibilities unfavorable to vilaprisan, choices of non-positive shift parameters $\delta_v = 0, -1, -2, \dots, \pi_{min}, \pi_{min} < 0$, for the log-odds of subjects with unclear response status in the vilaprisan group will be combined with choices of non-negative shift parameters $\delta_p = 0, 1, 2, \dots, \pi_{max}, \pi_{max} > 0$, for the log-odds of subjects with unclear response status in the placebo group. The combination $\delta_v = 0$ and $\delta_p = 0$ signifies a MAR assumption and represent the "best case" among the presented scenarios. In this scenario, patients with unclear response status are assumed to have the same chance of amenorrhea as similar patients for whom response status is known. Any combinations different from $\delta_v = 0$ and $\delta_p = 0$ refer to the situation where unobserved bleeding is missing not at random (MNAR), with larger absolute values for the penalities

indicating larger deviations from the MAR assumption. Values of δ_v<0 are equivalent to a multiplication of the odds (chance) of having amenorrhea with the factor $\exp(\delta_v)$ and imply vilaprisan patients with unclear response status have a systematically lower chance of amenorrhea than similar patients with clear response status. Choices of δ_p>0 or, equivalently, multiplication of the odds with $\exp(\delta_p)$ imply that placebo patients with unclear response status have a systematically higher chance of amenorrhea than those with clear response status. Therefore, a combination of $\delta_v = \pi_{min} < 0$ and $\delta_p = \pi_{max} > 0$ represents the "worst case" among these scenarios in which all subjects with unclear response status in the vilaprisan group are set to non-responders whereas all subjects with unclear response status in the placebo group are set to responders for amenorrhea.

6.2.2 Secondary efficacy variables

The secondary efficacy variables are defined in the following. The confirmatory analysis of selected secondary efficacy variables will be performed only for data from TP1. In addition, secondary efficacy variables will be analyzed descriptively for both treatment periods.

- **Heavy menstrual bleeding (HMB) response:** HMB response is defined as MBL <80.00 mL during last 28 days of treatment and >50% reduction from baseline (assessed by AH method). Missing AH data will be imputed as explained in Section 4.3.

  If an endometrial biopsy was conducted, bleeding on the day of intervention and the 3 days thereafter will not be considered in this evaluation, i.e., bleeding/'spotting' on the day of biopsy and 3 days thereafter will not be considered as day(s) with bleeding/'spotting' and AH values will be set to 0 mL.

  After imputing missing bleeding intensities and AH values (see Section 4.3) and considering potential biopsies, the 28-day period will be extended in case there are still any AH values missing in the last 28 days (similar to the description in Section 6.2.1 b); a) will not be applied). Based on the 28-day period a subject will be considered as not being an HMB responder if

  i) the sum of the non-missing (i.e., original or imputed) AH values is ≥80.00 mL *or* the reduction of blood loss is ≤50% as compared to baseline or

  ii) the subject did not complete at least 8 weeks of treatment.

  If neither of i) or iii) applies, the subject will be considered as an HMB responder.

- **Time to onset of amenorrhea**: Onset of amenorrhea is defined by the first day for which the MBL for all subsequent 28-day periods up to the end of a treatment period is <2 mL (amenorrhea defined similar to primary endpoint). The first 28 days with non-missing AH values until the last 28 days with non-missing AH values under treatment will be considered for deriving the time to onset of amenorrhea. An analogous approach as described for amenorrhea in the last 28 days of treatment (Section 6.2.1) will be applied to assess the amenorrhea status for each of the subsequent, overlapping 28-day periods within a treatment period (except for criterion iii), which does not apply). The censoring mechanism is assumed to be non-

informative and the subjects will be handled as right-censored, if applicable. If a subject completed the 84 days of a treatment period and did not experience an onset of amenorrhea, the subject will be censored on Day 57 of the treatment period.

For prematurely discontinued subjects, censoring also depends on whether an onset of amenorrhea can be ruled out or not during the time the subject was in the respective treatment period. An onset of amenorrhea can be ruled out if in the last 28-day period before the subject discontinued the respective treatment period, the sum of AH values exceeds 2 mL. If this is not the case, then an onset of amenorrhea cannot be ruled out for the respective treatment period. Thus, there are in principle the two possibilities:

- In case the subject discontinued from the respective treatment period prematurely and an onset can be ruled out, subjects will be treated as right-censored at the first day of the last 28-day period after which AH bleeding values for all further days are missing (see Figure 6–3 (a)).

If the subject prematurely discontinued from the respective treatment period but an onset cannot be ruled out, the subject will be censored on the day before the first day of the first 28-day period which indicates that an onset may have taken place (see Figure 6–3 (b)).

- **Time to onset of controlled bleeding**: Onset of controlled bleeding is defined by the first day for which the MBL for all subsequent 28-day periods up to the end of a treatment period is <80.00 mL. The first 28 days with non-missing AH values until the last 28 days with non-missing AH values under treatment in the respective treatment period will be considered for deriving the time to onset of controlled bleeding. An analogous approach as described for HMB response (Section 6.2.2) will be applied to each of the subsequent, overlapping 28-day periods within a treatment period (except for criterion ii), which does not apply). Time to onset of controlled bleeding is then derived analogous to time to onset of amenorrhea.

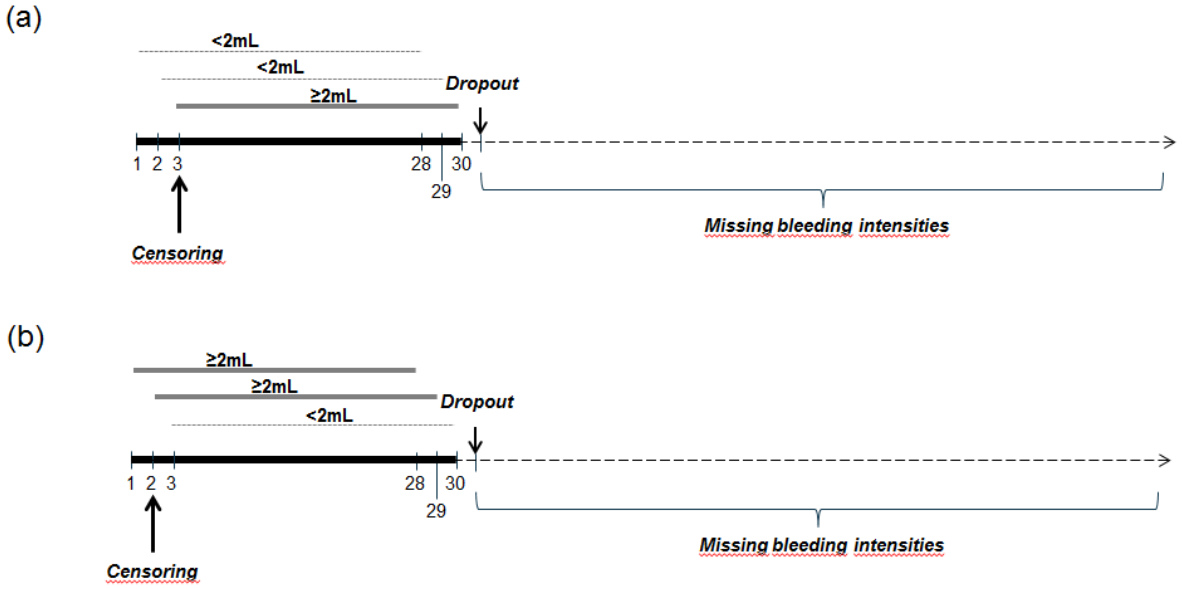
- **Absence of bleeding (spotting allowed) during last 28 days of treatment (assessed by the UF-DBD):** Absence of bleeding is defined as no scheduled or unscheduled bleeding during the last 28 days of a treatment period based on the UF-DBD. Missing data will be imputed as described in Section 4.3. The last days under treatment with 28 days with non-missing bleeding eDiary data will be used to calculate whether absence of bleeding occurred. Thus, if the last 28 days of treatment include days with missing bleeding eDiary data, the left margin of this interval will be displaced backwards in time and step-wise until the required number of 28 non-missing values are included.

  The left margin may be displaced backwards at maximum up to the end of the first bleeding episode when treatment was started in respective treatment period.

  If an endometrial biopsy was conducted, bleeding on the day of intervention and the 3 days thereafter will not be considered in this evaluation, i.e. bleeding on the day of biopsy and 3 days thereafter will not be considered as day(s) with bleeding and UF-DBD values will be set to ‘no vaginal bleeding’.

  Subjects who discontinue treatment prior to completion of 8 weeks of treatment will be considered as not experiencing absence of bleeding.

Figure 6–3: Illustration of censoring rules for subjects prematurely discontinuing a treatment period.



An illustration of censoring for a subject who both discontinued on Day 31 of a treatment period. (a) The onset of amenorrhea during the subject's treatment period can be ruled out (MBL ≥ 2mL in last 28-day period), and the subject is censored at the first day of her last 28-day period. (b) An onset of amenorrhea on Day 3 cannot be ruled out (MBL < 2mL in her last 28-day period(s)), and the subject is censored on Day 2.

6.2.2.1 Analysis of the secondary efficacy variables

HMB response and absence of bleeding (spotting allowed) will be analyzed descriptively in frequency tables by treatment period and region within each treatment group. The time to event variables will be analyzed for each treatment group using Kaplan-Meier estimates (including number of events, number of censored subjects, median, 25th and 75th percentiles). Kaplan-Meier plots will also be presented.

Additionally, treatment groups A1 and B2 with vilaprisan treatment in TP1 will be pooled and results will be presented for this pooled population.

The first three secondary endpoints evaluated for TP1 are part of the testing hierarchy within the confirmatory analysis. The confirmatory analysis of these variables is described in the following.

HMB response

The first secondary efficacy variable HMB response (yes/no) will be analyzed analogously to the primary efficacy variable amenorrhea.

Time to onset of amenorrhea

The time to onset of amenorrhea can be described in terms of a hazard rate. The null hypothesis of no treatment effect with respect to the onset of amenorrhea (see Figure 6–2) can thus be stated as an equality of hazard rates $\lambda_{v,k}(t)$ and $\lambda_{p,k}(t)$ in the k-th stratum/region for the vilaprisan group (v) and the placebo group (p):

$$H_0: \lambda_{v,k}(t) = \lambda_{p,k}(t) \text{ for } k = 1, 2, 3, t \leq \tau,$$

where k corresponds to the strata/regions (1 ≙ US, 2 ≙ Japan, and 3 ≙ other) and τ is the largest observed event time in the pooled treatment groups used for the comparison.

In order to test the null hypothesis of no difference in time to onset of amenorrhea between vilaprisan 2mg and placebo versus the alternative hypothesis of a difference for at least one stratum/region, a logrank test stratified by region/country is conducted at a local 0.05 significance level.

The test statistic of the logrank test compares the observed number of events and the numbers of events expected under the null hypothesis in each treatment group per stratum. It is defined as

$$Z = \sum_{k=1}^{3} Z_{v,k}(\tau) \Big/ \sqrt{\sum_{k=1}^{3} \hat{\sigma}_{v,k}^2},$$

where in each stratum k, the values for $Z_{v,k}$and $\hat{\sigma}_{v,k}$ are derived in the following way (the stratum index k is left out for all variables out of convenience):

$$Z_v(\tau) = \sum_{i=1}^{D} \left(d_{v,i} - Y_{v,i} \frac{d_i}{Y_i} \right) \quad ,$$

$$\hat{\sigma}_v^2(\tau) = \sum_{i=1}^{D} \frac{Y_{v,i}}{Y_i} \left(1 - \frac{Y_{v,i}}{Y_i} \right) \left(\frac{Y_i - d_i}{Y_i - 1} \right) d_i \ .$$

In these formulae, if $t_1 < \cdots < t_D = \tau$ are the distinct event times, i.e. times of onset of amenorrhea,

- the values of $d_{v,i}$ represent the number of events observed in the vilaprisan group at time point t_i
- $Y_{v,i}$ denotes the number of subjects at risk just prior to t_i in the vilaprisan group,
- d_i denotes the number of events observed in the vilaprisan group and placebo group combined at time point t_i and
- Y_i denotes the number of subjects at risk just prior to t_i in the vilaprisan group and placebo group combined [11].

The logrank test statistic Z asymptotically follows a standard normal distribution under the null hypothesis.

Time to onset of controlled bleeding

Time to onset of controlled bleeding will be analyzed analogously to the secondary efficacy variable time to onset of amenorrhea.

Absence of bleeding

Absence of bleeding (yes/no) will be analyzed descriptively with frequency tables by treatment period within each treatment group.

6.2.2.2 Sensitivity analyses of the secondary efficacy variables

Sensitivity analyses for the variables HMB response, time to onset of amenorrhea and time to onset of controlled bleeding will be conducted on the PPS analogously to the primary efficacy variable (see Section 6.2.1.2).

6.2.3 Further efficacy variables and analyses

All further variables will be presented by means of descriptive statistics based on the FAS population. Time to event variables will be analyzed using the Kaplan-Meier estimates (including number of events, number of censored subjects, median, 25th and 75th percentiles) and Kaplan-Meier plots will also be presented. For continuous variables, the number of observations, mean, median, standard deviation, minimum, and maximum as well as change from baseline will be computed. For categorical (dichotomous) variables the data will be presented using frequency tables. The data will be presented by treatment period and treatment group. Additionally, results for all patients receiving vilaprisan treatment for 12 weeks will be presented, i.e. results for TP1 of treatment groups A1 and B2 will be pooled.

The further efficacy variables and their analysis are outlined in the following subsections.

6.2.3.1 Binary bleeding variables

- **Treatment success rate**

  The treatment success rate is defined as percentage of subjects who fulfill the criteria of HMB response (MBL <80.00 mL and >50% reduction in MBL as compared to baseline) and who have not been withdrawn from the study due to AEs, due to non-fulfillment of selection criteria (related to fibroid size or HMB at baseline), due to fibroid surgery having been performed during the study. Treatment success rate will be derived separately for the AH method and for MP.

  Only fibroid surgeries that occur during the length of the originally planned study design (including FUP) prior to the temporary pause (Integrated Clinical Study Protocol, version 3.0, and protocol versions before) will be considered for deriving the treatment success rate. The FUP visit of the originally planned study design is on Day 7-15 of the 2nd menstrual cycle after the subject's end of treatment. That means that for subjects who have a prolonged FUP due to the temporary pause, the FUP according to the original study design will be derived. The derivation will be based on the end date of the subject's last treatment period and a follow-up time fixed to 12 weeks. Assuming a time to first menses after end of treatment of about 4 weeks and a cycle length of 4 weeks, 12 weeks of follow-up will cover at least the time period until the FUP visit would have been conducted according to the original study design.

  Treatment success will be summarized with frequency tables by treatment period within treatment group. Number of observations and percentages will be presented.

- **Amenorrhea**

  Amenorrhea, as further efficacy variable, is defined as MBL < 2 mL per 28 days of treatment for consecutive, non-overlapping 28-day periods not considered for the primary efficacy variable. An analogous approach as described for amenorrhea in the last 28 days of treatment (Section 6.2.1) will be applied to assess the amenorrhea status for each of the consecutive, non-overlapping 28-day periods within a treatment period (except for the 28-day period extension in case of missing values and criterion iii), which both do not apply).

  In case of no premature discontinuation, amenorrhea will be evaluated for the first and second 28-day periods of the subject's TP1 and TP2 when the AH method is used. When the MP is used, amenorrhea will be evaluated in all non-overlapping 28-day periods of TP1 and TP2. Only subjects with completed 28-day periods will be considered for the analysis of amenorrhea in those 28-day periods.

  In addition, amenorrhea will be analyzed for an incomplete period which does not comprise 28 days. This analysis is based on all subjects for whom the length of the treatment period is not a multiple of 28 days. The reason for this might be premature discontinuation from the study or prolonged treatment.

  Amenorrhea will be summarized by a frequency table by (pooled) treatment group and the 28-day periods grouped by treatment period.

- **HMB response**

  HMB response, as further efficacy variable, is defined as MBL <80.00 mL per 28 days of treatment and >50% reduction from baseline where MBL is assessed for consecutive, non-overlapping 28-day periods not considered for the secondary efficacy variable. An analogous approach as described for HMB response in the last 28 days of treatment (Section 6.2.2) will be applied to assess HMB response for each of the consecutive, non-overlapping 28-day periods within a treatment period (except for the 28-day period extension in case of missing values and criterion ii), which both do not apply).

  In case of no premature discontinuation, HMB response will be evaluated for the first and second 28-day periods of the subject's TP1 and TP2 when the AH method is used. When the MP is used, HMB response will be evaluated in all non-overlapping 28-day periods of TP1 and TP2. Only subjects with completed 28-day periods will be considered for the analysis of HMB response in those 28-day periods. In addition, HMB response will be analyzed for an incomplete period which does not comprise 28 days. This analysis is based on all subjects for whom the length of the treatment period is not a multiple of 28 days. The reason for this might be premature discontinuation from the study or prolonged treatment.

  HMB response will be summarized by a frequency table by (pooled) treatment group and the 28-day periods grouped by treatment period.

6.2.3.2 Time to onset variables

The time to onset variables described in the following will be analyzed analogous to the descriptive analysis described in Section 6.2.2.1, namely by Kaplan-Meier estimates and Kaplan-Meier plots for all regions combined.

- **Time to onset of amenorrhea** and **time to onset of controlled bleeding** (see Section 6.2.2)

  These variables will be assessed by the MP separately for TP1 and TP2.

- **Time to start of bleeding after last study drug intake**

  The start of bleeding after last study drug intake is defined by the first day for which bleeding intensity is "mild" or higher according to the UF-DBD. Only data from the UF-DBD will be considered to determine "start of bleeding after last study drug intake" because UF-DBD data are typically available even if there are missing values for one or both of the other two methods (MP, AH method).

  The last study drug intake refers to the end of TP1 and the end of TP2, as defined in Section 4.5 ("Treatment period").

  If an endometrial biopsy was conducted during the evaluation time period, any bleeding reported on the day of biopsy and the 3 days thereafter will not be considered in this evaluation, that is, if a bleeding intensity is reported on these days, the bleeding intensity will be overwritten to "no vaginal bleeding".

  If a subject did not experience a start of bleeding, she will be censored at latest available date from the bleeding eDiary data from the UF-DBD. If bleeding has to be induced, subjects will be censored on the date of bleeding induction. For the analyses after TP1, the subject will be censored at latest at TP2 start date if she starts TP2.

6.2.3.3 Volume of menstrual blood loss (MBL) related variables

The following variables related to volume of MBL will be evaluated both on the AH method and the MP:

- Volume of MBL per 28 days: Volume of MBL will be derived for each non-overlapping 28 days within a treatment period based on subjects who entered the respective 28-day period.
- Volume of MBL per bleeding episode for the first, second, and if applicable third bleeding episode after the end of treatment.
- Volume of MBL per bleeding episode within the treatment break. All days of the bleeding episodes will be used even if some days of the episode lie outside the treatment break. The first, second and third bleeding episode will be considered.
- Percentage of subjects with ≥ 50% reduction in MBL per each non-overlapping 28-day periods compared to baseline (based on subjects who entered the respective 28-day period).

Bleeding on the day of biopsy and the 3 days thereafter will not contribute to the MBL in this evaluation. Analysis of volume of MBL will be done by treatment group separately for TP1, break period, TP2 and FUP. An incomplete period which does not comprise 28 days will be analyzed for any of TP1, break period, TP2 and FUP. The analysis of this incomplete period

is based on all subjects for whom the length of the considered time period (TP1, break period, TP2 or FUP) is not a multiple of 28 days. A subject's data from the last $x < 28$ days of the considered time period will be included.

6.2.3.4 Ultrasound examination related variables

The following variables will be presented by (pooled) treatment group and by Intervals B as well as by post-baseline visit (i.e. Visit 4, Visit 5, Visit 6, EoT visit, FUP visit, safety closeout visit):

- Percent change in the sum of the volume of the 3 largest fibroids compared to baseline (measured by ultrasound). If less than 3 fibroids are present, the volume of the actual number of fibroids will be summed.
- Percentage change in volume of the largest fibroid from baseline (measured by ultrasound)
- Percent change in volume of uterus compared to baseline (measured by ultrasound)
- Percentage of subjects with a reduction of ≥ 25% of the sum of the volume of the 3 largest fibroids compared to baseline (measured by ultrasound). If less than 3 fibroids are present, the volume of the actual number of fibroids will be summed.
- Percentage of subjects with a volume reduction of ≥ 25% of the largest fibroid compared to baseline (measured by ultrasound)
- Percentage of subjects with a reduction of ≥ 25% of uterine volume compared to baseline (measured by ultrasound)

If multiple volume assessments are available for a time period, the smallest volume measurement will be used for the Baseline time period, and the largest volume measurement will be used for any other time periods of Intervals B.

6.2.3.5 Uterine fibroids surgeries

The number and proportions of subjects undergoing surgical treatment for uterine fibroids will be displayed in a frequency table by treatment group and by the following time periods: treatment period, break period, FUP (including prolonged FUP). In contrast to other efficacy variables, any fibroid surgeries performed during FUP of patients who did not start TP2 will be evaluated for FUP rather than for break period.

6.2.3.6 Patient-reported outcome and Clinician-reported outcome

Change in UF-DSD individual items compared to baseline

The UF-DSD is a newly developed multi-item PRO instrument designed to assess cardinal symptoms of uterine fibroids using a 24-hour recall period.

The UF-DSD includes 2 items (items 1 and 2) to assess swelling and bloating symptoms in the abdomen and pelvic area using a 5-graded Likert-type severity rating ('no symptom' to 'very severe symptom'), and 2 items (items 3 and 4) using a 0 to 10 numerical rating scale to assess pain at its worst in the abdominal/pelvic and lower back areas respectively, with 0 indicating 'no pain' and 10 'pain as bad as you can imagine'. Additionally, item 5 assess the

intake of pain medication using a 4-point verbal rating scale ("no"; "yes, over the counter (non-prescription) pain medication"; "yes, prescription pain medication"; or "yes, both (over the counter and prescription pain medication"). Further information can be found in Appendix 3).

The following variables in each domain will be summarized by treatment group. Mean values will be evaluated per 28-day period within each of the following time periods: screening, baseline (only the last 28 days before first study drug intake), 1st 28-day period, 2nd 28-day period, 3rd 28-day period for any of TP1, break, TP2, FUP1, FUP2, etc.. Absolute changes from baseline will be provided as well. Graphical displays will also be provided.

Bulk symptoms (items 1 and 2):

- Mean of subjects' relative frequencies (relative frequency is calculated for each subject as ratio of given severity to all severities) of the symptom scores
- Mean of subjects' symptom scores of the 7 days with worst assessment (i.e., highest score)

Pain (items 3 and 4):

- Mean numerical rating pain score
- Mean numerical rating pain score of the 7 days with worst pain (i.e., highest score)

Pain medication (item 5):

- Proportion of days with pain medication intake

Change in UFS-QoL scores compared to baseline

The UFS-QoL is a widely used disease-specific instrument that assesses symptom severity and health-related quality of life (HRQoL) in subjects with uterine fibroids [12].

It consists of 37 items including an 8-item symptom severity scale and 29 health-related quality of life questions comprising 6 subscales: concern, activities, energy/mood, control, self-consciousness, and sexual function which are responded to using a recall period of 3 months. All items are scored on a 5-point Likert scale, ranging from "not at all" to "a very great deal" for symptom severity items and "none of the time" to "all of the time" for the HRQoL items.

Symptom Severity, Activities, Revised Activities, HRQoL subscale scores and revised HRQoL scores are summed respectively and transformed into a 0-100-point scale each. The Symptom Severity scale and the HRQoL subscale scores are inversely related with higher Symptom Severity scores indicating greater symptoms while higher HRQoL subscale scores indicate better HRQoL.

Details can be found in Appendix 4.

For each subscale (including the symptom severity scale), raw scores and transformed scores will be summarized using descriptive statistics within treatment group by visit (Visit 3, Visit 5, EoT visit and FUP visit), as well as the absolute change from baseline. The time course for each subscale (raw scores and transformed scores) by visit and by treatment group will be shown using line plots depicting the mean ± standard deviation.

Change in CGI-I scores compared to baseline

The Clinical Global Impression (CGI-I) is a widely used single item questionnaire which asks the investigator to describe the subject's overall severity of uterine fibroids symptoms. The verbal response options for CGI-I are:

- None
- Very mild
- Mild
- Moderate
- Severe
- Very severe

Each item will be summarized by a frequency table with the number of observations and percentage for each treatment group. CGI-I is recorded at Visit 3 (baseline), EoT visit and FUP visit. CGI-I assessments will be shown by visit. Graphical displays of scores at the three visits will also be provided.

The change from baseline to EoT visit and FUP visit, respectively, will be analyzed using shift tables.

6.2.3.7 Variables related to laboratory parameters

Intervals C will be used for the following assessments:

- Change from baseline in hemoglobin, hematocrit, and ferritin:
  Absolute change from baseline will be summarized using descriptive statistics and displayed by time period within treatment group.
- Percentage of subjects with normal hemoglobin >12 g/dL and normal hematocrit >36% :
  Frequency tables with number of observations and percentage will be displayed by time period within treatment group.

If multiple measurements for a patient are available for any of the parameters during a time period, the smallest value is used in the analysis, as specified in Appendix 5.

6.2.4 Subgroup analyses

Descriptive statistics for the primary and the first three secondary efficacy variables will be reported separately for each region/country (the US, Japan, other countries), race and ethnicity for the FAS population.

6.3 Pharmacokinetics/pharmacodynamics

The planned analysis of pharmacokinetic and pharmacodynamics data will be described in a separate M&S Analysis Plan and results will be reported in a separate M&S Report.

6.4 Safety

The safety variables will include the following secondary and other safety variables:

The secondary safety variables are:

- Endometrial histology (e. g., benign endometrium, presence or absence of hyperplasia or malignancy)
- Endometrial thickness

The other safety variables are:

- Endometrial histology (diagnosis of Progesterone receptor modulator-associated changes (PAEC), individual features of PAEC)
- Ovarian cysts (number, size)
- Laboratory parameters
- AEs
- Cervical smear
- Vital signs
- Percentage of subjects with hemoglobin ≤10.9 g/ dL
- Findings resulting from liver monitoring
- Findings resulting from adrenal monitoring
- Findings resulting from skin monitoring
- UF-DBD bleeding pattern per 28 days and 84 days

All safety analyses are descriptive, and variables will be summarized by descriptive statistics as appropriate. All safety analyses are done for the actual treatment (as-treated) and performed on the SAF.

In contrast to the derivation of efficacy variables, even if an endometrial biopsy was conducted, bleeding and 'spotting' reported on such days will also be used when deriving safety variables.

No results for the pooled treatment group A1+B2 will be presented, if not specified otherwise.

Pregnancy test results based on ultrasound and urine pregnancy test will be reported by treatment group and visit. Ultrasound is performed as confirmation after positive urine pregnancy test.

6.4.1 Secondary safety variables

6.4.1.1 Endometrial histology

For endometrial histology evaluations (e.g., benign endometrium, presence or absence of hyperplasia or malignancy) the following time intervals / time points will be considered:

- Any time point

- Intervals A (acc to Figure 4–1)
- Intervals B (acc to Figure 4–1)
- "Exit examination" intervals
- Bleeding episode after last treatment period time windows in subjects with at least 56 days of treatment within the treatment period (by TP1, TP2 and overall):
    - From start of the patients' last treatment period until Day 55 of the same treatment period
    - From Day 56 of the same treatment period until the last day of the 1st bleeding episode after the treatment period
    - After 1st bleeding episode of the same treatment period, i.e., from 1st day after the end of the 1st bleeding episode to the last day of the 2nd bleeding episode
    - After 2nd bleeding episode of the same treatment period, i.e., from 1st day after the end of the 2nd bleeding episode to the last day of the 3rd bleeding episode
    - After 3rd bleeding episode of the same treatment period, i.e., from 1st day after the end of the 3rd bleeding episode to the last day of the 4th bleeding episode

Analyses will present the number and percentage of either subjects or biopsies. A subgroup analysis by exposure group (i.e., treatment duration ≤ 90 days, 91 to 180 days, etc.) will be presented for endometrial biopsy main results.

Presentations will be done for

- the Safety read (Reader #1 or #5)
- the Majority read (Reader #2 - #4)
- All reads (Reader #1 - #5)

Safety read: For the safety read only one result will be available per biopsy (based on either Reader #1 or #5), whereas for the "majority read" and "all reads" more than one reader result needs to be considered.

Majority read: Majority read results will only be provided in case biopsy data from Readers #2, #3 and #4 are available. Majority read will be determined for main results and subcategories (see Table 6–1). First, adequacy (for part II and III) or sufficiency (for part IV) of tissue will be investigated by all readers. If at least 2 of the 3 readers consider the tissue adequate/sufficient, the majority for the main results will be assessed. If there is a majority with respect to the main result, majority of the respective subcategories will be determined. If no majority result is available (3 different results in 3 readers, 2 different results in 2 readers), either "no consensus" or the worst case will be presented. Table 6–1 presents an overview of the biopsy results including the approach which is used in case no majority is available.

For subcategories where multiple selections are possible, an individual feature will be considered a majority read result if at least 2 readers have ticked this feature. That means that more than just one subcategory can be a majority read result for a biopsy.

All read: All read results will be based on all biopsies with results from at least one reader. A biopsy is considered non-benign in case "Benign endometrium" = "No" or "Endometrial Hyperplasia (WHO 2014)" or "Endometrial Hyperplasia (WHO 1994)" or "Malignant Neoplasm" = "Yes".

Multiple biopsies per time interval/time point: For the exit examination the last available measurement will be used if there are several measurements. For any other time intervals / time points, the analyses in which numbers of subjects are reported will be based on a 'worst case approach'. In this approach, the worst measurement within each of the above-mentioned time intervals / time points will be used. These worst outcomes across time intervals / time points are determined for each endpoint in Table 6–1.

Table 6–1: Overview of biopsy endpoints including majority result handling and worst case

Part		Endpoint	No majority available	Worst case across time intervals for subject-based analyses
Main results				
I	Main diagnosis	Adequate endometrial tissue	- *(not possible)*	"Yes"
II		Benign endometrium Endometrial Hyperplasia (WHO 2014) Malignant Neoplasm	Worst case: List is ordered by severity, from low to high	List is ordered by severity, from low to high
III		Endometrial Polyp	Worst case: yes	"Yes"
IV	PAEC	Tissue sufficient for PAEC read	- *(not possible)*	"Yes"
		PAEC present	Worst case: yes	"Yes"
Subcategories				
II	Main diagnosis	Benign endometrium (select one) ☐Atrophic ☐Inactive ☐Proliferative ☐Disordered Proliferative ☐Secretory including progestin and OCP effect ☐Menstrual ☐Endometritis ☐Other,specify	"no consensus"	- *(will not be presented)*
		Endometrial Hyperplasia (WHO 2014) (select one) ☐Hyperplasia without atypia ☐Atypical hyperplasia / Endometrioid Intraepithelial Neoplasia (EIN)	Worst case: Atypical hyperplasia / Endometrioid Intraepithelial Neoplasia (EIN)	Atypical hyperplasia / Endometrioid Intraepithelial Neoplasia (EIN)
III		Endometrial Polyp (select one) ☐Atrophic ☐Functional ☐Hyperplastic	"no consensus"	- *(will not be presented)*
IV	PAEC	PAEC (tick all that apply) ☐Pre-Decidua Absent ☐Extensive cysts Present ☐Secretory Changes, Extensive ☐Mitoses (required, may be rare) ☐Apoptosis ☐Abnormal Vessels	"no consensus"	- *(will not be presented)*

6.4.1.2 Endometrial thickness

For the analysis of endometrial thickness, the following time intervals will be considered:

- Intervals A (acc to Figure 4–1)
- Intervals C (acc to Figure 4–1)
- "Exit examination" intervals

Number of subjects with ultrasound performed for the time intervals will be presented as well.

Descriptive statistics for change from baseline in endometrial thickness will be provided. In addition, the proportion of subjects with endometrial thickness >18 mm will be investigated. A subgroup analysis by exposure group (i.e., treatment duration ≤ 90 days, 91 to 180 days,

etc.) will present the proportion of subjects with endometrial thickness >18 mm and cyst like structure either ovary > 3cm.

The correlation between endometrial thickness and biopsy results will be studied based on biopsies with endometrial thickness value measured on the day of an endometrial biopsy. The following proportions will be compared between patients with endometrial thickness >18 mm and patients with endometrial thickness ≤18 mm (measured by ultrasound):

- Proportion of biopsies with non-benign and benign results in all readers,
- Proportion of biopsies with PAEC present and PAEC not present,
- Proportion of biopsies with extensive cysts present and not present.

Intervals A will be used for this analysis.

6.4.2 Other safety variables

6.4.2.1 Endometrial histology (PAEC)

The analysis of progesterone receptor modulator-associated endometrial changes (PAEC) will be presented together with further results of endometrial histology. A detailed description of the analyses can be found in Section 6.4.1.1.

6.4.2.2 Ovarian cysts

Ovarian cysts will be examined by ultrasound. All scheduled and unscheduled assessments will be considered for the analysis. Three analyses will be performed for ovarian cysts:

- Number of subjects showing cyst like structures with largest diameter >3 cm in the ovary
- Analysis of ovarian cyst episodes (see below for definition)
- Analysis of AEs associated with ovarian cysts

The analysis of *Number of subjects with ovarian cyst like structures > 3 cm in largest diameter* in ovary will be shown by treatment group and by time intervals. Intervals A and Intervals B will be considered. One table will be produced for Intervals A and one table for Intervals B.

For each time interval it will be determined whether an individual subject shows cysts like structures with a largest diameter > 3 cm at least once during that time interval. In addition to the corresponding time intervals, both tables will contain the assessment for the ‘Exit examination’.

Frequencies for the type of cyst like structures identified will be included in the table. Type of cyst like structures will be shown as recorded in the eCRF: ‘Follicle-like structure’, ‘Corpus luteum cyst’, ‘Endometrioma’ and ‘Other’.

The overall assessment for both ovaries will be displayed, and information about ovary laterality will not be included in the table.

If cyst like structures with diameter > 3 cm are visualized in the ovaries, unscheduled ultrasound examinations should be performed at least every 4 weeks to document the regression or outcome. An *ovarian cyst episode* is defined as the time period from occurrence

of a cyst like structure with diameter > 3 cm until its resolution. Ovarian cyst episodes will only take account of the occurrence of follicle-like structures and corpus luteum cysts. Two different definitions of "resolution" are used in the analysis:

- The subject does not show in the same ovary (left or right) a cyst like structure of the same type,
- The subject does not show in the same ovary (left or right) a cyst like structure of the same type with diameter > 3 cm.

For both definitions, the date of the first observation of the cyst like structure is the *start of the episode*. The date when the cyst like structure appears for the first time as resolved (absence of cyst like structure or cyst like structure with diameter > 3 cm) is considered the *end of the episode*. The length of time between start and end of the episode is the *duration of the episode*.

If the cyst resolution does not occur, then the length of time between the start of the episode and the last observation will be considered the cyst episode duration.

If only the start of the episode is recorded but no follow-up has been performed, then the duration of the episode is unknown.

The analysis will include cysts episodes occurring in both ovaries with no distinction of ovary laterality (left or right). Although the information of ovary laterality will not be displayed in the table, it will be critical to monitor the progression of individual cyst like structures.

The Intervals A and B (acc to Figure 4–1) will be considered. Therefore, two sets of tables will be produced. Ovarian cyst episodes will be assigned to the time interval when the start of the episode occurs.

Two analyses will be produced for ovarian cyst episodes:

- Number of subjects with at least one ovarian cyst episode by time intervals.
- Number of ovarian cyst episodes by time intervals including: i) descriptive statistics of ovarian cyst episode duration, ii) by categorization of the duration by 4-week intervals, and iii) descriptive statistics of largest diameter of the cyst like structure per episode and per time interval.

Two tables will be produced for these two analyses, one for each definition of cyst episode resolution.

The analysis of AEs originating from ovarian cysts will include AEs identified by the Bayer MedDRA query (BMQ) for 'Benign ovarian cyst and associated complications' [V21], which include the preferred terms (PTs): Haemorrhagic ovarian cyst, Ovarian cyst, Ovarian cyst ruptured and Ovarian cyst torsion.

The table will show number of subjects with AEs, non-serious AEs and serious AEs (SAEs). For each one of these AE categories the table will include severity, whether the AE led to drug withdrawal (if applicable) and AE duration. In addition, SAEs will include the reason for being classified as serious as recorded in the eCRF.

The analysis will be done for pre-treatment AEs, TEAEs and post-treatment AEs. Pre-treatment and post-treatment AEs will be shown by treatment group (including pooled group A1+B2) only. TEAEs will be shown by treatment group for TP1, TP2 and overall, and additionally they will be shown by study drug. AEs with onset during a treatment period or up to 8 days after last study drug intake in any treatment period will be considered 'on-

treatment'. AEs with onset after date of last study drug intake in any treatment period + 8 days and before the start of the next treatment period will be considered 'off-treatment'. The summary table for TEAEs will be presented by overall AEs, on-treatment AEs and off-treatment AEs.

6.4.2.3 Laboratory parameters

Laboratory parameters which will be descriptively analyzed are included in Appendix 5. The tabulation of laboratory data will follow in principle the Bayer Global Standard Tables catalogue (currently version 4.0 default), with the necessary modifications like for time interval presentation instead of visit presentation. Treatment-emergent high and low laboratory abnormalities will be summarized by treatment periods (TP1, TP2) and overall.

For immunology and coagulation parameters Intervals 'BL/Post baseline' will be used, while for all other parameters Intervals C will be used (see Figure 4–1). If several laboratory measurements are available for a patient, depending on the laboratory parameter only either the smallest (if small values are considered worse for the parameter) or largest (if large values are considered worse) measurement within Intervals C will be used (see Appendix 5 for details). If both largest and smallest values are specified for a particular parameter, two separate analyses will be done for this parameter, one analysis using the patient's largest measurement across the considered time period and another analysis using the smallest measurement. Both scheduled and unscheduled measurements will be considered.

Central as well as local lab measurements are available. Descriptive statistics will be provided for central lab measurements. Local lab measurements will be listed.

Values below the lower limit of quantification (LLOQ) will be set to LLOQ/2. Values above the upper limit of quantification (ULOQ) will be set to ULOQ.

Estradiol

Absolute values will be analyzed using descriptive statistics by treatment group using Interval C and Interval D (see Figure 4–1). The smallest value (i.e. "worst case") per patient per time interval will be used in the analyses.

The distribution of estradiol values by treatment group using Intervals C and D will be graphically depicted with boxplots. The time course will be shown using line plots depicting the mean and 95% confidence interval (two-sided) and using Intervals C and D as described above. Time intervals including data from less than 10 patients in the respective group will be excluded from the plots.

For estradiol frequency tables with number and percentage of subjects with estradiol (pg/mL) <13, <30 will be displayed by treatment group and Intervals A and B. The subjects with at least one lower value based on each cut-off in any visit during that time interval will be counted.

6.4.2.4 Adverse events

All AEs will be coded by MedDRA terms (the current version at the time of analysis) and classified into pre-treatment AEs, treatment-emergent (TEAE) and post-treatment AEs.

AEs will be flagged as TEAE except for AEs for which there is clear evidence that the AE starts before date of first study drug intake (pre-treatment AEs) or after the date of last study drug intake + 60 days (post-treatment AEs).
AEs with clear evidence of start before the date of first study drug intake are flagged as pre-treatment AE unless the AE worsens at or after the date of first study drug intake and before or at the date of last study drug intake + 60 days. AEs starting before but worsening at or after the date of first study drug intake and before or at the date of last study drug intake + 60 days will be considered as two AEs, a pre-treatment AE and a TEAE. The pre-treatment AE will end at the date of worsening and the new TEAE will start at the date of worsening.

Same applies in case an AE starts at or after the date of first study drug intake, but before the date of last study drug intake + 60 days and worsens afterwards, then e.g. leading to a TEAE and a post-treatment AE.

Worsening of an AE is defined as follows:

- AE intensity/grade is worsened (e.g., moderate to severe)
- AE changed to a serious event
- AE ends with death
- AE is drug-related

TEAEs will further be categorized into on-treatment and off-treatment events.

TEAEs will be flagged as on-treatment except for TEAEs where there is clear evidence that the TEAE starts outside all on-treatment phases. Each on-treatment phase starts with the date of first study drug intake and ends with the date of last study drug intake during a treatment period plus 8 days. TEAEs with clear evidence of start outside all on-treatment phases will be flagged as off-treatment TEAEs unless the TEAE worsens during an on-treatment phase. TEAEs starting during off/on-treatment phase, but worsening during an on/off-treatment phase will be considered as two TEAEs, an off/on-treatment TEAE and an on/off-treatment TEAE. The off/on-treatment TEAE will end at the time of worsening and the on/off-treatment TEAE will start at the time of worsening.

Descriptive analysis will be performed (i) by treatment group for both TP1 and TP2 combined, (ii) by treatment group (including pooled treatment group A1+B2) separately for each treatment period and (iii) by study drug for both TP1 and TP2 combined. The tabulation will follow in principle the Bayer Global Standard Tables catalogue (currently V 4.0), with the necessary modifications like for on- and off-treatment TEAEs.

6.4.2.4.1 Assignment to study drug

Depending on the TEAE onset, the AE will be assigned to a study drug. This is relevant for subjects with switch in study drug (i.e., treatment groups B1 and B2), as it may not be clear whether the TEAE is assigned to the study drug given in TP1 or to the study drug given in TP2.

- AEs with an onset before the start of TP2 will be assigned to the study drug given in TP1.
- AEs with an onset at or after the start of TP2 will be assigned to the study drug taken during TP2.

Consequently, AEs with an onset before start of TP2 which are ongoing during TP2 will not be assigned to the study drug given in TP2 unless the TEAE worsens after start of TP2. In case it is not clear to which study drug the AE will be assigned (for example, due to partially missing dates), those AEs will be assigned to vilaprisan (conservative approach).

6.4.2.4.2 Classifications of adverse events

Serious Adverse Events (SAEs) will be summarized in the same way as described for TEAEs.

Non-serious AEs will be shown on a summary table by primary SOC/PT.

Adverse events of special interest (AESI) will be identified in two different ways:

1) The investigators will assess all AEs to determine if they are AEs of special interest (AESIs) and document this in the eCRF. The following AEs are included:
    a. HMB
    b. Liver disorders
    c. Endometrial disorders
    d. Skin disorders
    e. Adrenal disorders
2) Adverse events will be categorized into AE groupings of special interest based on MedDRA groupings as defined within expert statement for compound vilaprisan (Selection of coding conditions applicable for AESIs as defined in the study protocols of vilaprisan phase II/III studies).

Table 6–2: List of AESIs in current expert statement based on MedDRA Version 22.1 (version 5.0, status 19FEB2020)

No.	AESI	Grouping
1	HMB (especially after end of treatment): HMB will be documented in detail throughout the study	*MTG*: [BMQ] Increased female genital bleeding[V8]
2	Liver enzymes	*SMQ*: Drug related hepatic disorders - comprehensive search (SMQ)[V29]
3	Endometrial hyperplasia (all subcategories according to WHO 2014 (and WHO 1994) classification	*PT selection:* PT: Endometrial hyperplasia PT: Biopsy endometrium abnormal PT: Endometrial dysplasia PT: Endometrial metaplasia PT: Biopsy uterus abnormal
4	Endometrial thickening >18 mm	*PT selection:* PT: Endometrial thickening PT: Endometrial disorder

Table 6–2: List of AESIs in current expert statement based on MedDRA Version 22.1 (version 5.0, status 19FEB2020)

No.	AESI	Grouping
		PT: Endometrial hypertrophy
5	Adrenal neoplasms, Adrenal diagnoses related to cortical hyperfunction and cortical hypofunction	*PBMQ*: Adrenocortical gland pathology including tumors and related hormonal changes and conditions Vilaprisan (narrow scope) PT Addison's disease PT Adrenal adenoma PT Adrenal androgen deficiency PT Adrenal atrophy PT Adrenal calcification PT Adrenal cortex dysplasia PT Adrenal cortex necrosis PT Adrenal cyst PT Adrenal disorder PT Adrenal gland abscess PT Adrenal gland cancer PT Adrenal gland cancer metastatic PT Adrenal gland injury PT Adrenal gland tuberculosis PT Adrenal glomerular zone abnormal PT Adrenal haematoma PT Adrenal haemorrhage PT Adrenal insufficiency PT Adrenal insufficiency neonatal PT Adrenal mass PT Adrenal neoplasm PT Adrenal suppression PT Adrenal thrombosis PT Adrenalitis PT Adrenocortical carcinoma PT Adrenocortical insufficiency acute PT Adrenocortical insufficiency neonatal PT Adrenogenital syndrome PT Adrenoleukodystrophy PT Adrenomegaly PT Benign neoplasm of adrenal gland PT Catecholamine crisis PT Congenital adrenal gland hypoplasia PT Congenital anomaly of adrenal gland PT Cortisol deficiency PT Cushingoid PT Cushing's syndrome PT Familial glucocorticoid deficiency PT Glucocorticoid deficiency PT Haemorrhagic adrenal infarction PT Hyperadrenalism PT Hyperadrenocorticism PT Hyperaldosteronism PT Hypercorticoidism PT Hypoaldosteronism PT Metastases to adrenals PT Mineralocorticoid deficiency PT Myelolipoma

Table 6–2: List of AESIs in current expert statement based on MedDRA Version 22.1 (version 5.0, status 19FEB2020)

No.	AESI	Grouping
		PT Primary adrenal insufficiency PT Primary hyperaldosteronism PT Pseudohypoaldosteronism PT Secondary adrenocortical insufficiency PT Secondary aldosteronism PT Steroid withdrawal syndrome PT Triple A syndrome
6	Cutaneous sarcomas	*PT selection*: PT: Dermatofibrosarcoma protuberans PT: Atypical fibroxanthoma PT: Leiomyosarcoma PT: Liposarcoma PT: Skin angiosarcoma PT: Kaposi´s sarcoma PT: Malignant fibrous histiocytoma PT: Soft tissue sarcoma

AESIs based on investigator assessment and AESIs based on MedDRA groupings will be presented separately. They will be presented by AESI group (if available), MedDRA primary SOC and PT and study drug. Summary tables for AESIs by maximum intensity and by worst outcome will be produced. Furthermore, listings will be provided.

6.4.2.5 Cervical smear

A frequency table for cervical smears findings at screening (Visit 1), FUP visit and safety closeout visit will be produced by treatment group.

6.4.2.6 Vital signs

Vital signs (heart rate, systolic blood pressure and diastolic blood pressure, and weight) will be summarized by treatment group and visit, including change from baseline where appropriate, using descriptive statistics. Boxplot for absolute values at each visit will be presented by treatment group. The time course by visit and by treatment group will be shown using line plots depicting the mean and 95% confidence interval (two-sided).

At the safety closeout visit the blood pressure will be measured as triplicates. The mean value of those will be reported in tables.

6.4.2.7 Anemia

Frequency presentations will be provided for moderate to severe anemia with hemoglobin ≤10.9 g/dL. Frequency tables with number of subjects and percentage will be displayed within treatment group and by Intervals C. If a subject has more than one hemoglobin measurement within a time interval of Intervals C, the smallest value (i.e. worst) will be used.

A listing will be provided for subjects with abnormal values.

6.4.2.8 Findings resulting from liver monitoring

Hepatic safety will be evaluated based on Intervals A and BL/Post baseline. The following parameters will be investigated in addition to the standard lab presentations:

- Aspartate aminotransferase (AST) (in U/L),
- Alanine aminotransferase (ALT) (in U/L),
- Alkaline phosphatase (AP) (in U/L),
- Bilirubin in serum (in mg/dL).

If a subject has more than one measurement of a parameter within the above specified time interval, the largest value (i.e. worst) will be used for the statistical analysis.

Frequency tables presenting number and percentage of subjects will be presented for the following categorizations:

- For ALT and AST, separately:
  - Normal (≤1×ULN); >1×ULN to <3×ULN; ≥3×ULN to <5×ULN; ≥5×ULN to < 8×ULN; ≥8×ULN to <10×ULN; ≥10×ULN to <20×ULN; ≥20×ULN
  - ≥ 3×ULN, ≥ 5×ULN, ≥ 8×ULN, ≥ 10×ULN, ≥ 20×ULN
- For ALT and AST combined (if at least one of ALT and AST falls into the category):

  ≥ 3×ULN, ≥ 5×ULN, ≥ 8×ULN, ≥ 10×ULN, ≥ 20×ULN
- For bilirubin:
  - >1×ULN, >2×ULN
  - ≤1×ULN; >1×ULN to ≤2×ULN; >2×ULN
- For AP:
  - >1.5×ULN
  - ≤1×ULN; >1×ULN to ≤1.5×ULN, >1.5×ULN
- For combinations of ALT/AST and bilirubin:
  - >3×ULN for ALT or AST accompanied by >1.5×ULN in bilirubin (measured at any time point)
  - >3×ULN for ALT or AST accompanied by >2×ULN in bilirubin (measured at any time point)
  - ≥3×ULN of ALT or AST accompanied by ≥2×ULN of bilirubin (measured within 30 days afterwards) (Hy's Law criteria). The event will be assigned to the time interval of the ALT/AST elevation.

Kaplan Meier estimates for the time to ALT > 3×ULN will be derived for each considered time interval. If no such an increase is observed within the considered time interval, the observation is censored at the end date of the time interval. Tables with the number of

subjects under risk, cumulative number of subjects with ALT > 3×ULN, and estimated probability for an event including 95% confidence intervals (two-sided) will be presented per 28 days. Furthermore, Kaplan Meier plots will be provided.

eDISH plots (x-axis: maximum of ALT or AST, resp., within time interval; y-axis: maximum of total bilirubin within 30 days thereafter) and a scatter plot (x-axis: maximum ALT within time interval; y-axis: maximum AST within time interval) will be provided for each considered time interval.

If a patient has ALT ≥ 3×ULN at any time point, a plot for her individual time course in the following laboratory parameters will be presented: ALT, AST, bilirubin, and AP relative to ULN over time. It will be indicated within the plot on which days the study drug was taken (i.e. start and stop of treatment periods). Furthermore, listings will be provided with results relative to ULN for liver-related parameters, i.e. ALT, AST, bilirubin, AP.

6.4.2.9 Liver symptom inquiry

Before the implementation of Integrated Clinical Study Protocol, version 5.0, investigators were asked to regularly inquire about symptoms that according to their medical judgement may indicate liver disturbance and document the result of this inquiry in the respective eCRF page. This data will be listed only.

6.4.2.10 Findings resulting from adrenal monitoring

Subjects were asked to undergo scheduled adrenal monitoring at the safety closeout visit. The adrenal monitoring comprises the following assessments:

- Adrenal gland imaging
- Laboratory testing
- Signs and symptoms suggestive of hypercortisolism, hyperaldosteronism

All tables generated within this adrenal monitoring section will contain the absolute number of subjects and percentage among subjects who performed the safety closeout visit.

An *overview* table will be provided. This table includes the number and percentage of subjects taking part at the specific adrenal monitoring assessments. The following numbers and percentages of subjects will be provided in addition:

- Subjects with abnormal result in at least one of the assessments.
- Subjects with abnormal result for each specific assessment separately.
- Subjects for whom an external expert recommendation was obtained. Among those, also the number and percentage of subjects with any findings suspicious of an adrenal disorder and recommendation for further handling will be provided.
- Subjects who have seen a local specialist (including information on type of specialist).

Details of the *final diagnosis by the expert panel* will be provided in another table. This table contains the number and percentage of subjects assessed by an expert, subjects needed expert panel meeting, subjects where expert panel met, and subjects with at least one diagnosis by the expert panel. Among those, the following numbers and percentages of subjects will be provided:

- Subjects with agreement of the diagnosis among all three experts.
- Subjects with a specific diagnosis (multiple diagnoses are possible).

A listing of subjects with other diagnosis (i.e. diagnosis entered by the expert panel in a text field) will be provided. A listing of subjects with adrenal tumor will presented as well. It includes diagnosis, time from last study drug until date of panel assessment and possible causal relationship to study drug. Cases where the three experts did not agree on a diagnosis, are presented in a separate listing including also the diagnosis of the local specialist.

A third table will summarize the *diagnosis by a local specialist*. This table contains the number and percentage of subjects with at least one diagnosis by the local specialist. Among those, the numbers and percentages of subjects with a specific diagnosis (multiple diagnoses are possible) will be reported.

A listing of subjects with other diagnosis (i.e. diagnosis entered by a local expert in a text field) will be provided as well as a listing of reasons if a local specialist consultation did not take place despite referral.

A listing of local laboratory measurements with reference ranges will be provided (no descriptive analyses will be performed on those).

A table with *results of the adrenal glands imaging* will be provided. This table contains the number and percentage of subjects who did and who did not perform adrenal imaging (and reasons if not performed), and who have at least one evaluable image. Among subjects having performed adrenal glands imaging the type of procedure (e.g. computed tomography [CT] scan with contrast) will be reported. The following numbers and percentages of subjects will be derived among subjects with at least one evaluable image:

- Subjects with one or two evaluable lateralities.
- Subjects with adrenal mass (if yes, in one or in both adrenal glands).
- Subjects with specified diagnostic benign imaging features.
- Subjects with specified indeterminate imaging features.

Reader results where an adjudicator was needed due to discrepant diagnoses between the two readers will be provided in a listing.

6.4.2.11 Findings resulting from skin monitoring

Findings resulting from skin monitoring will be analyzed descriptively by treatment group. All subjects who consent for safety follow-up and attend the safety closeout visit will undergo a thorough skin examination by a dermatology expert. The information will be recorded on the eCRF assigned to the safety closeout visit. The analysis will show the number of subjects who consented to participate in the Safety Follow-up, how many were assessed by a dermatologist or other clinician, which diagnostic tests were used and the results of the assessment as follows: number of normal and abnormal findings; and following an abnormal finding which type of abnormality as recorded in the eCRF: ‘Precancerous skin lesion’, ‘Cutaneous sarcoma’, ‘Other malignant skin tumor’ and ‘Other skin abnormality’.

Relevant cases will be reported as AEs, potentially as an AESI (pre-cancerous and malignant skin lesions). Malignant skin tumors will be classified as SAEs. Narratives will be written for skin-related AESI and SAEs.

6.4.2.12 Bleeding patterns per 28 days and per 84 days based on UF-DBD

The bleeding data recorded on the UF-DBD will be summarized using descriptive statistics in the SAF by treatment groups (including pooled treatment group A1+B2) using Intervals C and D. A reference period of 28 days will be used for most analyses, which corresponds to the Intervals D. When a reference period of 84 days is used, the Intervals D will be modified accordingly.

The bleeding intensity will be summarized by bleeding category using frequency tables by reference period of 28 days and 84 days.

The maximum bleeding intensity will be summarized using frequency tables by 28 days reference period.

The number of bleeding days will be analyzed by 28 days reference period and with the following bleeding categories:

- non-bleeding,
- spotting-only,
- bleeding (bleeding intensity of 'mild' or higher),
- bleeding including spotting (bleeding intensity of 'spotting' or higher),
- mild,
- moderate,
- severe,
- very severe.

An incomplete period which does not comprise 28 (or 84, respectively) days will be analyzed for any of the considered time periods of Intervals C. The analysis of this incomplete period is based on all subjects for whom the length of the considered time period is not a multiple of 28 (or 84, respectively) days. A subject's data from the last $x < 28$ ($x < 84$) days of the considered time period will be included.

7. Document history and changes in the planned statistical analysis

The following variables will be analyzed descriptively, even those were not listed as other efficacy/safety variables in the protocol:

- Percentage change in volume of the largest fibroid from baseline (measured by ultrasound)
- Percentage of subjects with a volume reduction of ≥ 25% of the largest fibroid compared to baseline (measured by ultrasound)
- Uterine fibroid surgery performed during the study
- Bleeding patterns per 28 days and per 84 days based on UF-DBD

8. References

[1] Best practice document RD-SOP-1107: Recording and Evaluation of Bleeding Data. Version 10.0

[2] Zivković N, Zivković K, Despot A, Paić J, Zelić A. Measuring the volume of uterine fibroids using 2-and 3-dimensional ultrasound and comparison with histopathology. Acta Clin Croat. 2012 Dec 1;51(4):579-89.

[3] Upton GJ. Conditional independence, the Mantel-Haenszel test, and the Yates correction. The American Statistician. 2000 May 1;54(2):112-5.

[4] Agresti A. Categorical data analysis. 2nd edition. Hoboken, New Jersey: John Wiley & Sons; 2003.

[5] Greenland S, Robins JM. Estimation of a common effect parameter from sparse follow-up data. Biometrics. 1985 Mar 1:55-68.

[6] Sato T. On the variance estimator for the Mantel-Haenszel risk difference. Biometrics. 1989;45(4):1323-4.

[7] National Research Council. The prevention and treatment of missing data in clinical trials. Washington, DC: National Academy Press; 2010.

[8] Carpenter J, Kenward M. Multiple imputation and its application. 1st edition, Chichester, UK, John Wiley & Sons; 2012.

[9] SAS/STAT 13.1 User's Guide. Cary, NC: SAS Institute Inc.; 2013. 5033p.

[10] Rubin DB. Multiple imputation for nonresponse in surveys. New York: Wiley; 1987.

[11] Klein JP, Moeschberger ML. Survival analysis: techniques for censored and truncated data. New York, Springer; 2003.

[12] Spies JB, Coyne K, Guaou NG, Boyle D, Skyrnarz-Murphy K, Gonzalves SM. The UFS-QOL, a new disease-specific symptom and health-related quality of life questionnaire for leiomyomata. Obstetrics & Gynecology. 2002 Feb 1;99(2):290-300.

9. Appendices

Appendix 1: Menstrual pictogram (MP) mapping to menstrual blood loss volume (in mL)

During the study, subjects will be required to use selected types of sanitary products (pads and/or tampons), and to assess the intensity of their menstrual blood loss per sanitary product using a visual scoring system. The analysis value can be obtained from Table 9–1.

Table 9–1: Mapping of the menstrual pictogram (MP) to menstrual blood loss

Sanitary product	Menstrual spot shape	Menstrual blood loss (mL)
always Ultra normal	Towel pictogram a	0.79
	Towel pictogram b	1.51
	Towel pictogram c	3.08
	Towel pictogram d	4.64
	Towel pictogram e	6.75
	Towel pictogram f	11.91
always Ultra night	Towel pictogram a	0.74
	Towel pictogram b	2.20
	Towel pictogram c	4.21
	Towel pictogram d	6.84
	Towel pictogram e	9.42
	Towel pictogram f	15.2
always Ultra long	Towel pictogram a	1.36
	Towel pictogram b	1.89
	Towel pictogram c	3.68
	Towel pictogram d	6.12
	Towel pictogram e	8.69
	Towel pictogram f	13.41
o.b. Pro Comfort light	Tampon pictogram a	0.16
	Tampon pictogram b	0.55
	Tampon pictogram c	1.19
	Tampon pictogram d	2.98
o.b. Pro Comfort regular	Tampon pictogram a	0.35
	Tampon pictogram b	1.31
	Tampon pictogram c	2.09
	Tampon pictogram d	5.66
o.b. Pro Comfort super	Tampon pictogram a	1.29
	Tampon pictogram b	1.87
	Tampon pictogram c	3.49
	Tampon pictogram d	9.37
o.b. Pro Comfort super plus	Tampon pictogram a	1.34
	Tampon pictogram b	3.36
	Tampon pictogram c	5.81
	Tampon pictogram d	12.23
Tampax light	Tampon pictogram a	1.24
	Tampon pictogram b	1.52
	Tampon pictogram c	1.73
	Tampon pictogram d	3.10
Tampax regular	Tampon pictogram a	0.63
	Tampon pictogram b	1.58
	Tampon pictogram c	2.20
	Tampon pictogram d	5.62

Table 9–1: Mapping of the menstrual pictogram (MP) to menstrual blood loss

Sanitary product	Menstrual spot shape	Menstrual blood loss (mL)
Tampax super	Tampon pictogram a	1.29
	Tampon pictogram b	2.38
	Tampon pictogram c	4.01
	Tampon pictogram d	7.76
Tampax super plus	Tampon pictogram a	1.20
	Tampon pictogram b	2.56
	Tampon pictogram c	4.98
	Tampon pictogram d	9.49
Other Tampon	Tampon pictogram a	3.55
	Tampon pictogram b	4.33
	Tampon pictogram c	6.25
	Tampon pictogram d	8.30

Appendix 2: Uterine Fibroid Daily Bleeding Diary (UF-DBD)

Subjects will be asked to rate any vaginal bleeding in the past 24 hours on a daily basis in the eDiary. The respective question was: "Rate the severity of any vaginal bleeding in the past 24 hours". The following answers could be selected in the eDiary by the subjects:

- ☐ No vaginal bleeding
- ☐ Spotting
- ☐ Mild
- ☐ Moderate
- ☐ Severe
- ☐ Very severe

Appendix 3: Uterine Fibroid Daily Symptoms Diary (UF-DSD)

Table 9–2: Uterine Fibroid Daily Symptoms Diary questionnaire

Item	Response scoring
1. Rate the severity of any swelling in your abdomen or pelvic area in the past 24 hours	5 point Likert scale
2. Rate the severity of any bloating in your abdomen or pelvic area in the past 24 hours	5 point Likert scale
3. How severe was the pain in your abdomen or pelvic area, at its WORST, in the past 24 hours?	10 point scale
4. How severe was the pain in your lower back, at its WORST, in the past 24 hours?	10 point scale
5. Have you taken any pain medication in the past 24 hours?	No Yes, over the counter Yes, prescription Yes, both

Categories Item 1:

- No swelling
- Mild
- Moderate
- Severe
- Very severe

Categories Item 2:

- No bloating
- Mild
- Moderate
- Severe
- Very severe

Appendix 4: Uterine Fibroid Symptom and Quality of Life questionnaire (UFS-QoL)

To calculate a symptom score for the *symptom severity*, a summed score based on items 1-8 is created. This will provide symptom scores with higher score values indicative of greater symptom severity or bother, and lower scores are indicative of lesser symptom severity.

Based on the lowest possible score that can be obtained and the range of raw scores, an observed sum score for the symptom severity can be projected into a percentage according to the transformation:

$$\textit{Transformed symptom severity score} = \left[\frac{(Actual\ raw\ score - Lowest\ possible\ raw\ score\ )}{Possible\ raw\ score\ range} \times 100\%\right],$$

with transformed scores taking values in the range 0% to 100%.

The lowest possible raw score for symptom severity as well as the score range is provided in Table 9–3.

Table 9–3: Description of the UFS-QoL scoring system for assessing symptom severity.

Scale	Sum item values	Lowest and highest possible raw scores	Possible raw score range
Symptom severity	Sum scores of items 1-8	8, 40	32

Similar to the symptom severity, raw and transformed scores for the *Health-Related Quality of Life (HRQoL) subscales* can be derived. The HRQoL subscales comprise the 6 domains: concern, activities, energy/mood, control, self-conscious, and sexual function. The sum score for the HRQoL subscales can be derived based on the items 9-37 (see Table 9–4). Based on

the highest possible score and the raw score range, these raw sum scores can be transformed to a percentage according to the transformation:

$$Transformed\ HRQoL\ score = \left[\frac{(Highest\ possible\ score - Actual\ raw\ score)}{Possible\ raw\ score\ range} \times 100\%\right].$$

The respective highest possible raw scores and the range for each subscale is provided in Table 9–4.

Note that the scale is inverted when transforming raw HRQoL scores, such that lower *raw* HRQoL scores indicate a better HRQoL, while lower *transformed* HRQoL scores indicate a worse HRQoL.

Table 9–4: Description of the UFS-QoL scoring system.

Scale	Sum item values	Lowest and highest possible raw scores	Possible raw score range
Concern	I9+I15+I22+I28+I32	5, 25	20
Activities	I10+I11+I13+I19+I20+I27+I29	7, 35	28
Energy/mood	I12+I17+I23+I24+I25+I31+I35	7, 35	28
Control	I14+I16+I26+I30+I34	5, 25	20
Self-conscious	I18+I21+I33	3, 15	12
Sexual function	I36+I37	2, 10	8
HRQoL total score	Sum of 6 subscale scores	29,145	116
Revised activities	I11+I13+I19+I20+I27	5, 25	20
Revised HRQoL total score	Sum of 6 subscale scores (replacing activities with revised activities)	27, 135	108

Item scores for items 9 to 37 are abbreviated with I9 to I37.

To calculate the (raw) HRQoL total score, the values of the individual subscales are summed (note that this is not the sum of the individual items). Transformed HRQoL total scores based on the formula above can be obtained as well.

For the subscale analyses, if < 50% of items of a subscale is missing, the subscale should be retained with the mean subscale score generated based on the non-missing items ("half item rule"). If ≥ 50% of the items are missing, no subscale score should be calculated, and the subscale score should be considered missing. If one or more subscale scores are missing, the HRQoL total score is not calculated.

Appendix 5: List of laboratory parameters

Table 9–5: List of laboratory parameters

		In case of multiple measurements for a patient, which measurement to use	
Laboratory category	Laboratory test	Smallest	Largest
VITAMINS	25-Hydroxyvitamin D	X	
COAGULATION	Activated Partial Thromboplastin Time		X
HORMONES	Adrenocorticotropic Hormone	X	X
GENERAL CHEMISTRY	Alanine Aminotransferase		X
GENERAL CHEMISTRY	Albumin	X	
GENERAL CHEMISTRY	Alkaline Phosphatase		X
IMMUNOLOGY	Alpha-1 Antitrypsin		X
IMMUNOLOGY	Anti Mitochondiral Antibody		X
IMMUNOLOGY	Antinuclear Antibodies		X
GENERAL CHEMISTRY	Aspartate Aminotransferase		X
HEMATOLOGY	Basophils	X	X
HEMATOLOGY	Basophils/Leukocytes		X
GENERAL CHEMISTRY	Bilirubin		X
URINALYSIS	Bilirubin		X
URINALYSIS	Blood		X
GENERAL CHEMISTRY	Bone Specific Alkaline Phosphatase		X
MICROBIOLOGY	Brucellae		X
GENERAL CHEMISTRY	Calcium	X	X
PROTEINS	Ceruloplasmin		X
GENERAL CHEMISTRY	Chloride		X
GENERAL CHEMISTRY	Cholesterol		X
	Choriogonadotropin Beta		X
HORMONES	Choriogonadotropin Beta		X
URINALYSIS	Choriogonadotropin Beta		X
HORMONES	Cortisol	X	X
GENERAL CHEMISTRY	Creatine Kinase		X
GENERAL CHEMISTRY	Creatinine		X
IMMUNOLOGY	Cytomegalovirus DNA		X
GENERAL CHEMISTRY	Direct Bilirubin		X
HEMATOLOGY	Eosinophils		X
HEMATOLOGY	Eosinophils/Leukocytes		X
IMMUNOLOGY	Epstein-Barr Capsid IgG Antibody		X
IMMUNOLOGY	Epstein-Barr Capsid IgM Antibody		X
IMMUNOLOGY	Epstein-Barr Nuclear Antigen 1 IgG Ab		X
HEMATOLOGY	Ery. Mean Corpuscular Hemoglobin		X
HEMATOLOGY	Ery. Mean Corpuscular Volume	X	
HEMATOLOGY	Erythrocytes	X	
HORMONES	Estradiol	X	
GENERAL CHEMISTRY	Ferritin	X	
HORMONES	Follicle Stimulating Hormone		X
GENERAL CHEMISTRY	Gamma Glutamyl Transferase		X
URINALYSIS	Glucose		X
IMMUNOLOGY	HCV PCR Viral Load		X
GENERAL CHEMISTRY	HDL Cholesterol	X	X
HEMATOLOGY	Hematocrit	X	
HEMATOLOGY	Hemoglobin	X	
GENERAL CHEMISTRY	Hemoglobin A1C		X
IMMUNOLOGY	Hepatitis A Virus Antibody		X
IMMUNOLOGY	Hepatitis A Virus Antibody IgM		X
IMMUNOLOGY	Hepatitis B Virus Core Antibody		X
IMMUNOLOGY	Hepatitis B Virus DNA		X
IMMUNOLOGY	Hepatitis B Virus Surface Antibody		X
IMMUNOLOGY	Hepatitis B Virus Surface Antigen		X
IMMUNOLOGY	Hepatitis C Virus Antibody Surface		X
IMMUNOLOGY	Hepatitis D Virus Antibody		X
IMMUNOLOGY	Hepatitis D Virus IgG Antibody		X
IMMUNOLOGY	Hepatitis D Virus IgM Antibody		X
IMMUNOLOGY	Hepatitis D Virus RNA		X
IMMUNOLOGY	Hepatitis E Virus Antibody		X
IMMUNOLOGY	Hepatitis E Virus IgG Antibody		X

Table 9–5: List of laboratory parameters

Laboratory category	Laboratory test	In case of multiple measurements for a patient, which measurement to use	
		Smallest	Largest
IMMUNOLOGY	Hepatitis E Virus IgM Antibody		X
IMMUNOLOGY	Hepatitis E Virus RNA		X
IMMUNOLOGY	Herpes Simplex Virus 1 IgG Antibody		X
IMMUNOLOGY	Herpes Simplex Virus 1 IgM Antibody		X
IMMUNOLOGY	Immunoglobulin A		X
IMMUNOLOGY	Immunoglobulin G		X
IMMUNOLOGY	Immunoglobulin M		X
GENERAL CHEMISTRY	Iron	X	
URINALYSIS	Ketones		X
GENERAL CHEMISTRY	Lactate Dehydrogenase		X
GENERAL CHEMISTRY	LDL Cholesterol		X
MICROBIOLOGY	Leptospirae		X
HEMATOLOGY	Leukocytes	X	
URINALYSIS	Leukocytes		X
HORMONES	Luteinizing Hormone		X
HEMATOLOGY	Lymphocytes	X	
HEMATOLOGY	Lymphocytes/Leukocytes		X
HEMATOLOGY	Monocytes	X	
HEMATOLOGY	Monocytes/Leukocytes		X
IMMUNOLOGY	MPO ANCA		X
HEMATOLOGY	Neutrophils	X	
HEMATOLOGY	Neutrophils/Leukocytes		X
URINALYSIS	Nitrite		X
GENERAL CHEMISTRY	Osteocalcin		X
HEMATOLOGY	Platelets	X	
GENERAL CHEMISTRY	Potassium	X	x
IMMUNOLOGY	PR3 ANCA		X
HORMONES	Progesterone	X	
HORMONES	Prolactin		X
GENERAL CHEMISTRY	Protein		X
URINALYSIS	Protein		X
COAGULATION	Prothrombin Intl. Normalized Ratio	X	
COAGULATION	Prothrombin Time		X
GENERAL CHEMISTRY	Pseudocholinesterase		X
IMMUNOLOGY	Smooth Muscle Antibody		X
GENERAL CHEMISTRY	Sodium	X	
HORMONES	Testosterone		X
HORMONES	Thyrotropin	X	X
GENERAL CHEMISTRY	Time to event		X
GENERAL CHEMISTRY	Total Iron Binding Capacity		X
IMMUNOLOGY	Toxoplasma gondii antibody		X
GENERAL CHEMISTRY	Triglycerides		X
GENERAL CHEMISTRY	Type I Collagen C-Telopeptides		X
URINALYSIS	Urobilinogen		X
	all parameters not otherwise specified		X