Document Type:	Statistical Analysis Plan
Official Title:	A randomized, double-blind, open for active comparator, parallel-group, multicenter Phase 2b study to assess the efficacy and safety of three different doses of P2X3 antagonist (BAY 1817080) versus placebo and elagolix 150 mg in women with symptomatic endometriosis
NCT Number:	NCT04614246
Document Date:	01-JUL-2022

Title page

A randomized, double-blind, open for active comparator, parallel-group, multicenter Phase 2b study to assess the efficacy and safety of three different doses of P2X3 antagonist (BAY 1817080) versus placebo and elagolix 150 mg in women with symptomatic endometriosis

Assess safety and efficacy of three different doses of P2X3 antagonist in women with symptomatic endometriosis

Bayer study drug	BAY 1817080		
Study purpose:	dose-finding		
Clinical study phase:	IIb	**Date:**	01 JUL 2022
Study No.:	20584	**Version:**	3
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
ALT	Alanine-aminotransferase
AP	Alkaline Phosphatase
AST	Aspartate aminotransferase
BEP	Biomarker Evaluation Plan
CI	Confidence Interval
COI	Concept of Interest
CSR	Clinical Study Report
EAPP	Endometriosis-Associated Pelvic Pain
EIS	Endometriosis Impact Scale
eDiary	electronic diary
ESD	Endometriosis Symptom Diary
EQ-5D-5L	European Quality of Life 5 Dimension 5 Level Scale
FAS	Full Analysis Set
pFAS	Primary Full Analysis Set
FUP	Follow-Up Period
HEOR	Health Economics and Outcome Research
HRQoL	Health-related Quality of Life
IE	Intercurrent Event
IQR	Inter-quartile range
LOS	Listing only set
MAH	Meaningful Aspect of Health
MCP-Mod	Multiple Comparison Procedures – Modelling
MedDRA	Medical Dictionary for Regulatory Activities
MMRM	Mixed Model Repeated Measurement
mWPAI	modified Work Productivity and Activities Impairment
N/A	Not applicable
NMPP	Nonmenstrual pelvic pain
NRS	Numerical rating scale
P2X3	Purinergic receptor P2X
PCS	Pain Catastrophizing Score
PGI-C	Patient Global Impression of Change
PGI-S	Patient Global Impression of Severity
PK	Pharmacokinetic(s)
PPS	Per-protocol set
PRO	Patient report outcome
pPPS	Primary Per-protocol set
SAE	Serious adverse event
SAF	Safety analysis set
SAP	Statistical Analysis Plan
SARS-CoV-2	Severe Acute Respiratory Syndrome Coronavirus 2
SAS	Statistical Analysis System
SFU	Safety Follow-Up
STD	Standard Deviation
TBL	total bilirubin
TEAE(s)	Treatment emergent adverse event(s)
ULN	upper limit of normal (all referring to serum/plasma)
VAS	Visual Analogue Scale
WHO-DD	WHO Drug Dictionary

1. Introduction

Endometriosis is a hormone-dependent gynecological disease characterized by painful symptoms like dysmenorrhea, chronic and non-cyclic pelvic pain, and dyspareunia. Considering all available treatment modalities, recent data confirm a significant degree of unmet medical need, since up to 70 % of treated patients have persistent symptoms like chronic pelvic pain that are not sufficiently managed (Bayer market research, 2009; 21,700 women in 8 countries).

Pathologically the disease is defined by the presence of endometrial tissue outside the uterine cavity (endometriotic lesions). The endometrial lesions exhibit signs of inflammation. BAY 1817080 demonstrated robust efficacy in different in vivo and in vitro models including effects on neurogenic inflammation with high relevance for endometriosis. BAY 1817080 is an antagonist for the P2X3 receptor that is an important mediator of nociception and other disorders with an over-activation of sensory nerve fibers and can be found in nerve fibers of endometriotic lesions. Thus, BAY 1817080 is believed to carry the potential to relieve endometriosis-associated pelvic pain and improve quality of life of women with endometriosis.

This Statistical Analysis Plan (SAP) is based on:

- the clinical study protocol BAY 1817080 / 20584, version 3.0, amendment number 2 dated 27 JAN 2022.

2. Study Objectives

The objective of this study was to identify the optimal dose of P2X3 receptor antagonist BAY 1817080 in women with symptomatic endometriosis and further assess efficacy and characterize safety and tolerability profile of BAY 1817080.

The **primary** study objective was to

- assess the dose-response relationship and demonstrate efficacy of BAY 1817080 compared to placebo in women with symptomatic endometriosis

The **secondary** study objectives were to

- identify at least 1 superior effective dose of BAY 1817080 compared to placebo
- evaluate the safety and tolerability of 3 doses of BAY 1817080 compared to placebo and elagolix 150mg in women with symptomatic endometriosis

Other objectives were to

- assess efficacy of BAY 1817080 compared to placebo and elagolix 150mg in the treatment of EAPP at weeks 4 / 8 and end of intervention period
- assess efficacy of BAY 1817080 compared to placebo and elagolix 150mg in the treatment of EAPP during days with vaginal bleeding (dysmenorrhea/menstrual pelvic pain) at weeks 4/ 8 and end of intervention period
- assess efficacy of BAY 1817080 compared to placebo and elagolix 150mg in the treatment of EAPP during days without vaginal bleeding (‘non-menstrual pelvic pain’) at weeks 4 / 8 / and of intervention period
- further describe the patient population, efficacy profile of BAY 1817080 and change during the study period using the data collected by patient reported outcomes
- assess sustainability of treatment effect and recurrence of symptoms

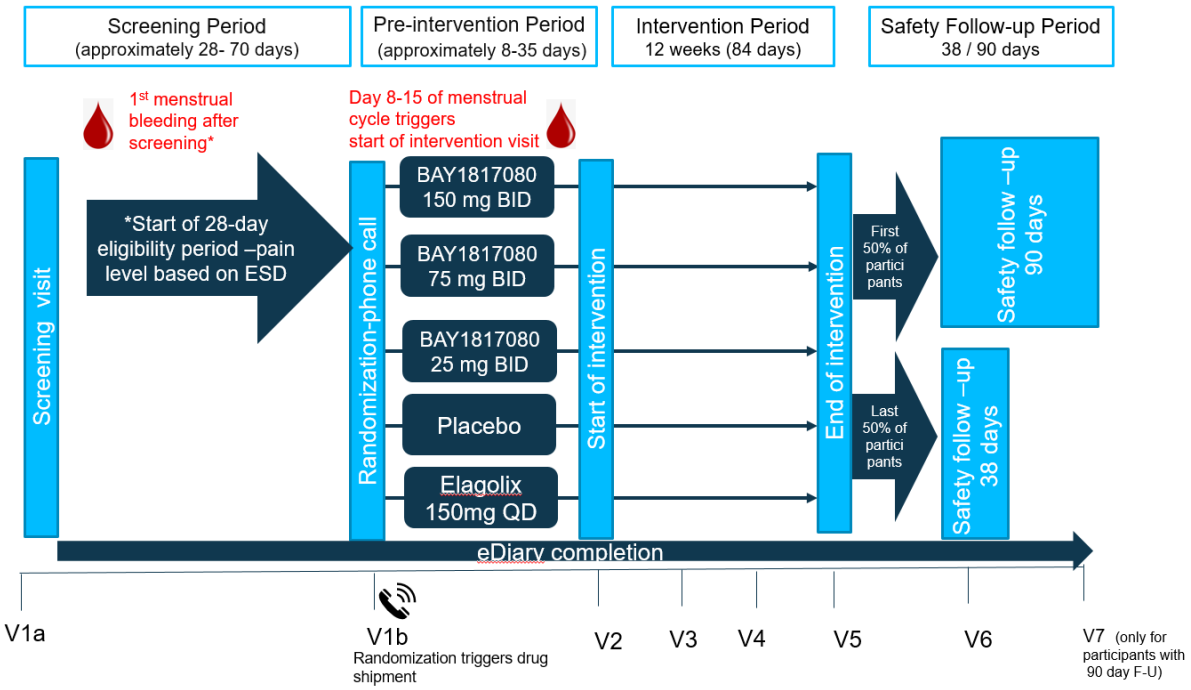
- describe sleep and physical activity associated measurements collected with accelerometer

A table that lists all primary, secondary, and exploratory objectives with the corresponding endpoints can be found in section 3 of the CSP. Additional other objectives listed in the CSP refer to analyses that are outside the scope of this SAP, and thus are not described in this document.

3. Study Design

This study is a placebo and active-controlled, double-blind, but open-label for the active comparator, randomized, parallel-group multi-center study. The study will be conducted in multiple clinical centers across Europe, North America, China, and Japan who take care of women with symptomatic endometriosis. An overview of the study design is shown in Figure 1.

Figure 1: Schema



*Follow-up 2 will be performed additionally in the 1st 50 % of randomized participants only. Follow-up 2 will extend to 90±3 days after the last administration of study intervention

Five treatment groups will be investigated: 25 mg, 75 mg, and 150 mg BAY 1817080 (referred to as low, mid, and high dosage group), placebo group and Elagolix group. The subjects will take 4 tablets twice daily (thus, 8 per day), containing BAY 1817080 or matching placebo, or 1 capsule once daily, containing 150 mg Elagolix. The treatment duration will be 12 weeks (84+3 days).

To achieve 50 evaluable participants for the primary analysis in each treatment arm, 840 participants were planned to be screened to start study intervention in about 420 randomized participants. Among them approximately 28 subjects from Japan with clinically confirmed instead of visually confirmed endometriosis and approximately 28 participants from China were planned to be randomised. None of the two groups will be included in the primary analysis.

The Japanese women with clinical confirmation will not be included as they are deviating from the overall population who is included with a surgically confirmed diagnosis. Women from China are mainly included to generate further data using the ESD in this population.

All participants that were found to be eligible according to the eligibility criteria assessable at screening visit (Visit 1a) will enter the screening period. Eligibility criteria related to the endometriosis symptom diary (ESD) will be calculated on the basis of the 28 consecutive days starting on the first day with menstrual bleeding (according to the definition given in Section 4.5) at or after Visit 1a. Thus, the screening period will last 28 to 70 days to allow for slightly irregular menstrual cycles and menstrual bleeding started immediately before the screening visit[1].

Re-screening is allowed once and only under the conditions stated in the CSP. Repeating the 28-day period for pain evaluation by the ESD is not explicitly excluded, because of many possible special cases that could justify a repetition. Not meeting the pain criteria per se provides no justification. Re-screened participant will be assigned a new participant number.

Eligible subjects will be randomized at Visit 1b. Participants from China and Japan are not randomized to the Elagolix group. The randomization ratio for the other treatment groups is set to 1:1:1:1 in China and Japan. A randomization ratio outside of China and Japan of 3:3:3:3:4 (Placebo:25 mg:75 mg:150 mg:elagolix) yields similar sizes of all treatment groups overall.

The follow-up period will last for 38±3 days and 90±3 days after the last dose of study intervention for all of the randomized participants, respectively.

3.1 Sample Size Determination

Approximately 840 participants will be screened to start study intervention in about 420 randomized participants for at least 50 participants evaluable for the primary analysis (i.e. included in the pPPS) per intervention group. The primary analysis will not include the participants enrolled based on visually-confirmed diagnosis, i.e. about half of the women randomized in Japan, nor all the participants randomized in China, who are mainly included to generate further data using the ESD in this population. Refer to Section 9.3 for further information of inclusion of participants into analysis sets.

If blinded data reviews indicate an unexpected high number of screening failures due to the COVID-19 pandemic, enrollment may be increased.

Sample size calculations were performed for establishing evidence of a drug effect across the doses, that is, detecting a statistically significant dose response signal for the primary efficacy outcome in this study using the MCP-Mod approach. They are based on the following assumptions:

- change in mean EAPP under placebo after 12 weeks of -2.3
- maximum change in mean EAPP of -1.2 for BAY 1817080 over placebo
- standard deviation of 2.5 for all dose groups
- a set of plausible dose-response shapes including Emax and sigmoidal Emax models

[1] The investigators are advised to schedule the screening visit during the second half of the menstrual cycle to keep the screening period as short as possible.

(chosen based on literature review and historic in-house data)

- random allocation of participants to dose groups according to a 1:1:1:1 ratio,

A sample size of 50 participants per dose groups will have appr. 90% power (averaged across the set of dose-response shapes) to demonstrate a dose- response relationship, using a onesided test at a type I error rate of α=0.10. Approximately 70 participants for primary analysis will be randomized to each treatment group to achieve 50 evaluable participants for each treatment group.

In addition, 28 participants from Japan with clinical diagnosis, as well as 28 participants from China will be randomized 1:1:1:1, but will not be included in the primary analysis. Although not formally included in the sample size determination, the exploratory elagolix 150mg treatment group is planned to be of about the same size as the other treatment groups. As participants from Japan and China cannot be randomized to the elagolix 150mg arm, the randomization ratio outside of Japan and China will be 3:3:3:3:4 (Placebo:25 mg:75 mg:150 mg:elagolix). This will provide similar sizes of all treatment groups overall.

4. General Statistical Considerations

4.1 General Principles

The statistical evaluation will be performed by using the software SAS (release 9.4 or higher; SAS Institute Inc., Cary, NC, USA) and ValidR (version 3.5.2 or higher; Mango Solutions Ltd., UK)[1].

All variables will be analyzed by descriptive statistical methods. The number of data available and missing data, arithmetic mean, standard deviation, minimum, quartiles, median, and maximum will be calculated for metric data. Frequency tables will be generated for categorical variables. Where appropriate, the individual change from baseline will also be analyzed.

Unscheduled/extra assessments (e.g. laboratory data or vital signs) associated with non-protocol visits will be included in listings, but not in any visit-based summary tables.

In general, confidence intervals (CIs) will be two-sided with a confidence level of 95 %. For the primary analysis of the primary efficacy variable only, 80% CIs will be reported in addition for consistency with the one-sided α-level of 10%.

4.2 Handling of Dropouts

A randomized participant who 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 participants who prematurely discontinue the study and study intervention for any reason, as well as the reasons for premature discontinuation of study and study intervention will be reported.

All dropouts will be carefully evaluated with respect to baseline characteristics. If necessary, additional sensitivity analyses may be performed.

The handling of missing data is described in section 4.3.

4.3 Handling of Missing Data

All missing or partial data will be presented in the subject data listings as they are recorded in the Case Report Form (CRF), electronic diary (eDiary) or device.

Subjects with missing baseline assessments will be excluded from all analyses that require the respective baseline assessment.

If the treatment end date is missing from the eCRF, the last documented exposure date in the ESD e-Diary will be used instead.

4.3.1 COVID-19 pandemic-related missing data

Efficacy Variables

Data collected during an active COVID-19 infection (confirmed by a positive SARS-CoV-2 test by PCR or an AE confirming COVID-19 infection) will be used in the primary efficacy analysis as they are not expected to be biased.

Treatment interruptions and additional pandemic-related reasons for treatment interruption will be collected and the time of the interruption will be subtracted from the exposure duration (for details see Section 6.4.2).

The number of participants who prematurely discontinue the study and study intervention for any pandemic-related reason, as well as the reasons for premature discontinuation of study and study intervention will be reported.

The number of subjects with missed visits and the number of subjects with missing e-diary data due to COVID-19 pandemic will be displayed as well as the reason why the visit/ assessment was not performed.

4.3.2 Diary data

28-day mean of ESD single items on pain

According to the ESD manual, the 28-day mean of an item asking for pain can be derived if at least 15 non-missing entries[2] are available (i. e., the number of missing entries is < 50 %).

For calculating the primary efficacy variable only, missing entries on days during a severe COVID-19 infection as recorded on the adverse event (AE) case report form (CRF) (this means in particular that the severity is not missing) will be imputed as described in Section 6.2.1 first. Thus, the 28-day will be derived only if the number of observed and imputed daily values add up to 15 or more.

Vaginal bleeding intensity

Missing bleeding diary data will be imputed following the sponsor's instruction on the evaluation of bleeding data.

4-week mean of Endometriosis Impact Scale (EIS) single items

A 4-week average score for a respondent will only be calculated if there are at least 2 not necessarily consecutive weeks of non-missing PRO data. This is consistent with the

[2] By design of the eDiary (ESD and EIS), if one item is missing, all items are missing. The respondents have to select an answer for an item in order to move on to the next item.

requirement of (at least) 50 % of the data to be available for calculating the 28-day mean of the ESD single items on pain.

4.3.3 Partially missing start/stop dates

Partially missing start dates/stop dates allow for the true date lying within a certain time window that is restricted by the respective stop date/start date and the reporting date. The partially missing date/s will be imputed by the middle of the related time window in which the true date is located.

For partially/completely missing start or end dates of prior/ concomitant medication, a worst-case assumption is made, i.e. if there is the possibility that the medication was taken during the intervention period, it will be classified as “concomitant”.

This worst case assumption will be used analogously for imputing partially missing start dates and stop dates of AEs. AEs will be classified as treatment-emergent, if possible. On that basis, the duration of an AE being classified as severe will be set as long as possible. Thereafter, the duration of the classification as moderate will be maximized. Missing intensities will not be imputed.

For calculating the primary variable only, the duration of a COVID-19 infection being classified as severe will be kept to a minimum of at least 1 day to avoid unjustified imputations (see Section 4.3.2).

4.4 Interim Analyses and Data Monitoring

Two interim analysis were planned during the course of the study.

The first interim analyses was to be performed on blinded data and after approximately 50% of the study participants had completed the 12-week intervention period in order to further validate the PRO instruments used in this study.

The second interim analysis was to be performed on unblinded data once the last participant had started treatment. These interim results were to be used for planning of Phase 3 studies. No final dose decision were to be made at this point in time.

Neither of the interim analyses were planned for stopping for efficacy or futility.

The final analysis for the study will occur at the conclusion of the study. No Type-I error adjustment will be made.

As the study was terminated early, none of these interim analyses will be carried out.

4.5 Data Rules

Definition of Baseline

In general, baseline is defined as the last available measurement before start of treatment.

The 28-day baseline period for the ESD items is the last 28 days before start of treatment (e.g. -1 to -28). In case these contain less than 15 available entries, the baseline value will be given by the last 28 consecutive days before the start of treatment with at least 15 non-missing values (e.g. -2 to -29, -3 to -30, -4 to -31 and so on).

In addition, for daily measurements, if baseline and screening periods overlap, all the overlapping measurements will be considered as baseline data.

Definition of Post Baseline

In general, post baseline is defined as the first available measurement after start of treatment. Intervention period is 84 days, capturing 3 menstrual cycles of 28 days on average. As defined in Section 4.3.2, the 28-day mean of an item asking for pain can be derived if at least 15 non-missing entries are available.

Therefore, a post baseline value will be given by 28 consecutive days starting from end of treatment and counting backwards (e.g. day 84 to day 57 will be cycle 3). If a participant does not have 3 full available cycles, the count will start from the last available full cycle with at least 15 non-missing entries available (e.g. if two cycles are available, the count will start from day 56 to 29).

If a participant has 87 days of measurements, the first 3 measurements on day 1,2 and 3 will be discarded.

In addition, if a participant has more than 87 days of measurement, each data collected after day 87 will be discarded.

(Absolute change from) Baseline

To calculate the absolute change from baseline, the baseline value will be subtracted from the post-baseline value of interest, i. e.,

$$\text{Absolute change} = \text{post baseline value} - \text{baseline value}.$$

PROs data rules

If the weekly EIS questionnaire is administered less than 3 days after starting either Week 1, Week 5, 9 or 13 after start of treatment, the collected questionnaires will be counted for the previous 28-day cycle.

For example if a participant has the questionnaire pushed on day 3 after starting her treatment period, the responses will be counted as baseline.

If the monthly EQ5D questionnaire is administered less than 3 days after starting either Week 1, Week 5, 9 or 13 after start of treatment, the collected questionnaires will be counted for the previous 28-day cycle.

For example if a participant has the EQ5D questionnaire pushed on day 3 after starting her treatment period, the responses will be counted as baseline.

The above rule for the EQ5D will also apply to the follow-up cycles.

Bleeding episode

The sponsor's standard procedures defines a bleeding episode as "day(s) with bleeding/spotting of which at least one day is of intensity "light" or higher, preceded and followed by at least 2 bleed-free days" [2] . Bleeding episodes will be derived from the ESD.

An episode is left censored if it starts at or before day 2 of the diary. Left censored episodes are expected to count to the previous cycle / reference period and are therefore ignored in all

episode related evaluations. Bleeding / spotting days of left censored episodes are evaluated in all day related evaluations as they would for not left censored episodes.

Likewise, for bleeding/spotting on the second last or last study day, the unobservable following two bleed-free days will be assumed to be given.

The reason for this is that each bleeding/spotting episode, spotting only episode, or bleeding free interval is attributed to the reference period in which it begins.

In addition, if a participant does not experience any bleeding days during the 28-day cycle, the observations collected will not be assessed for the dysmenorrhea endpoint.

Numerical coding of ESD values when response is not expected by design

The ESD device is programmed such that items 2 (worst constant pain) and 3 (worst short-term pain) are not administered if the participant reports no pain for item 1 (pain at its worst). Based on the logical relation of the questions, the values for items 2 and 3 will be set to 0 when item 1 = 0.

COVID-19 infection

The start of an active COVID-19 infection is defined as the date of the first positive SARS-CoV-2 RNA test or the start date of the AE confirming the COVID-19 infection, whichever is earliest. For participants with documented vaccination against COVID-19 during the study, positive serology IgG tests post vaccination will not be considered an active COVID-19 infection.

For the relevant MedDRA preferred terms see Appendix 3 [3]

The distinguishment of P2X3 related taste AEs from COVID-19 related smell/taste symptoms will be done by individual patient review based on the details of smell/taste disturbance assessments, the SARS-CoV-2/serology IgG test results, the AEs confirming COVID-19 and the timing information.

Rescue Medication

Ibuprofen rescue medication will be reported as 400mg. For the country in which this is collected as 200mg (Japan and USA), the number of tablets collected will be divided by 2.

Repeated measurements at the same visit

If more than one measurement flagged as valid for analysis is available for a given visit, the valid measurement closest to the start of treatment will be used in the data summaries. This is the last valid measurement before and the first valid measurement after the start of treatment visit. In the data listings, all observations will be presented.

Data Rules for laboratory measurements

For lab values reported as “<x”, “<x.x”,“ <x.xx” or etc. the value for analysis used will be derived by “x/2”, “x.x/2”,“ x.xx/2” etc . For lab values reported as “>x”, “>x.x”,“ >x.xx” or etc. the value for analysis used will be derived by increasing the last digit by 1.

4.6 Blind Review

The purpose of the blind review is to provide documentation of important deviations from the protocol and validity findings and the resulting assignment of subjects to the analysis sets (see section 5.1). This will be done according to the sponsor's applicable operational instruction [4][2]. The definition for important deviations and validity findings will be set down in the specification of assessment criteria and identification requirements before the blind review's closure.

Any changes to the statistical analysis prompted by the results of the review of study data will be documented in an amendment and, if applicable, in a supplement to this SAP before unblinding the data.

The blind review includes the continued monitoring of the number of participants who contracted COVID-19 (see Section 4.5), those otherwise affected by the pandemic (i. e., pandemic- related administrative reasons), and the related amount of missing data relevant for the primary analysis, including premature discontinuations of the study (treament). The impact of the amount of missing data on the power of the primary analysis may be assessed by running simulations. Patient characteristics of participants with pandemic-related missing data may be compared with those of participants with observed data. The latter two analyses will not be reported.

5. Analysis Sets

5.1 Assignment of analysis sets

Final decisions regarding the assignment of participants to analysis sets will be made during the blinded 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).

Enrolled

All participants who sign the Informed consent form (ICF).

Full analysis set (FAS)

All participants randomly assigned to study intervention. Participants will be analyzed according to the intervention they were randomized to.

Safety analysis set (SAF)

All participants randomly assigned to study intervention and who take at least 1 tablet of study intervention. Participants will be analyzed according to the intervention they actually received. If participants received different interventions throughout the course of the study, it will be decided in the assessment meeting on a case-by-case decision which group they will be assigned to for the final analysis. These participants will be excluded from the unblinded interim analysis.

Primary full analysis set (pFAS)

Subset of FAS, containing all participants with surgically confirmed endometriosis and outside China randomly assigned to study intervention. Participants will be analyzed according to the intervention they were randomized to.

Per protocol set (PPS)

All participants randomly assigned to study intervention, who take at least 1 tablet of study intervention, and who have no validity findings affecting efficacy.

A list of potential validity findings will be provided in a separate important deviations and validity specifications document (Assessment Criteria and Identification Requirement (ACIR)) which will be finalized before unblinding. The assignment of participants to this analysis set will be based on the assessment meetings.

The participants with missing baseline ESD item 1 (as defined in 4.5) or no valid post-baseline 28-day average of ESD item 1 will be excluded from the PPS.

Due to the repeated measurements design of the study a participant dropping out of the study may still be evaluable for efficacy and will therefore not be automatically excluded from PPS.

Participants will be analyzed according to the intervention they actually received. If participants received different interventions throughout the course of the study, they will be excluded from the PPS.

Primary Per Protocol Set (pPPS)

Subset of the PPS containing all participants in the PPS with surgically confirmed endometriosis and outside China. The analyses performed on the pPPS are chosen to be the primary one.

The PPS and pPPS will be used for all non-safety analyses and the separate pharmacodynamics analyses.

Pharmacokinetic Analysis Set (PKS)

All participants who took at least 1 dose of study drug and had at least 1 available pharmacokinetic sample.

Listing Only Set (LOS)

This set contains all screening failures. It will be used for listing purposes only.

The primary and secondary efficacy endpoints will be analyzed using the pPPS. The safety endpoints will be analyzed using the SAF. All other endpoints will be analyzed on the FAS, unless specified otherwise.

digitalBIO Analysis Set (DAS)

All participants who consented to take part in the actigraphy sub-study with any amount of actigraphy data available will be included in the Digital Measures Analysis Set. Participant may be included or excluded from particular analyses depending on the extent of available data. Participants will be analyzed according to the intervention they were randomized to.

6. Statistical Methodology

6.1 Population characteristics

All variables in this section will be analyzed descriptively. The analyses will be based on the FAS and pPPS, unless otherwise specified.

Any table displaying demographics and baseline characteristics will be repeated for the following COVID-19 pandemic related subgroups:

all randomized subjects affected by COVID-19 pandemic related study disruption

all randomized subjects with COVID-19 adverse event and/or positive SARS-CoV-2 PCR test and/or positive serology IgG test after start of treatment

6.1.1 Subject disposition and validity

All except one table in this section will be based on all enrolled subjects.

For each investigator, the number of screening failures and the number of protocol deviations and prematurely discontinuations of randomized subjects will be presented in total and per individual treatment group. The screening failures, discontinued subjects and protocol deviations will be listed along with the reasons, where applicable, each. In particular, the number of participants who prematurely discontinue the study for any COVID-19 pandemic-related reason, as well as the reasons for premature discontinuation of study will be reported. An additional listing will show the inclusion and exclusion criteria not met.

Participants who completed the study will be summarized. A participant completed the study if she completed all phases of the study including the last visit.

In addition, a disposition summary by study period (epoch), i.e. screening, treatment, and follow-up, will show the number of subjects completing the respective epoch and the number of subjects discontinuing it prematurely including information of the reasons for discontinuation. The table will include COVID-19 pandemic associated reasons for discontinuation, i.e. the information whether decision for discontinuation was made by the participant, the physician or was due to logistical reasons. An overview of the subject disposition including the number of participants who received at least one dose of study drug will be given overall and by treatment group.

For each investigator, the number of randomized participants valid for the SAF, the FAS, the PPS and the PKS will be presented in total and per treatment group.

An overview of the subject validity and the primary reasons for exclusion from the respective analysis set will be given overall and by treatment group for the FAS only. A listing of the subjects' assignment to the (p)FAS, (p)PPS, SAF and PKS and the reasons for exclusion will also be provided by treatment group.

6.1.2 Demographics and Baseline Characteristics

The following demographic data and baseline characteristics will be summarized:

Age at baseline (years);

Age category (<30years vs 30-40years vs >40-55years);

Race;

Ethnicity;

Height (cm);

Subject rescreened;

Natural hair color (at age 18);

Main eye color;

Weight (kg) at baseline;

BMI (kg/m²) at baseline;

Reproductive and menstrual history;

Endometriosis history/characteristics

- Age at first endometriosis treatment (medical or surgical whatever is earlier);
- Previous medical treatments;
- Number of surgeries;

Baseline pain

- Mean worst pain;
- Nonmenstrual pelvic pain (NMPP);
- Dysmenorrhea;

Bleeding characteristics;

Baseline scores of EIS.

Demographic variables and baseline characteristics will be summarized using descriptive statistics and will be reported for the FAS and PPS.

6.1.3 Medical history

Medical history findings will be coded according to the latest effective version of the Medical Dictionary for Regulatory Activities (MedDRA), Medical history findings (id est, previous diagnoses, diseases or surgeries starting before signing of the informed consent) considered relevant to the study will be summarized by primary system organ class (SOC) and preferred term overall and per treatment group. The table will be calculated based on the SAF and the FAS, respectively.

6.1.4 Prior and concomitant medication

The number of participants that used prior and concomitant medication will be analyzed using frequency tables based on classified data. The classification will be done according to the World Health Organization Drug Dictionary (WHO-DD) and Bayer Drug Grouping (BDG).

This includes a classification of the substances which will be used for presentation, such as medication prohibited due to potential drug-drug interaction (E.g. Analgesia producing opiods and Nonsteroidal anti-inflamatory drugs (NSAIDs).

The tables will display the following three categories: prior, prior/concomitant and post-treatment.

Regarding medication data collected during the study course, the following rules will be implemented:

If the stop time of medication is before the time of the first study drug administration, then medication is considered to be **prior** medication.

If the start time of medication is before or at the time of the first study drug administration and the stop time of medication is after or at the time of the first study drug administration then the medication is considered to be **concomitant** medication.

If the start time of medication is at or after the time of the first study drug administration and earlier than 14 days after or at the time of the last study drug administration, then the medication is considered to be **concomitant** medication.

If the start time of medication is later than 14 days after the time of the last study drug administration, then the medication is considered to be **post-treatment** medication.

If time information is missing but only date information is given, then classification of medication as prior or concomitant will be done by date information only. In such case, if the start day of medication is the same date as the first study drug administration, then the medication will be classified as concomitant medication.

Analysis of prior and concomitant medication will be performed on the SAF. Prior, concomitant and post-treatment medication will be shown by treatment group and total.

6.2 Efficacy

6.2.1 Primary Efficacy Endpoint

The primary efficacy endpoint is defined below along with other attributes of the primary estimand.

Primary Estimand

Endpoint: The primary efficacy variable will be the absolute change in mean worst EAPP from baseline (last 28 days before the first intake of study drug) to end of intervention (last 28 days ending with the last intake of study drug planned on Day 84 (+3)). The worst EAPP will be measured daily on the NRS by item 1 of the ESD. The time frame of 28 days captures a menstrual cycle on average. For more details on how to derive the mean values over 28 days see Section 4.5.

Summary measure: A population-level treatment group summary is an estimated model-based group mean of the primary endpoint's values adjusted for region (Japan vs. ROW excluding China) and a baseline EAPP. Treatment effect will be evaluated as a difference between group means.

Treatment: The investigated treatments will be BAY 1817080 25 mg or 75 mg or 150 mg or placebo, plus standardized rescue medication taken for any reason.

Population: The target population is comprised of women with surgically confirmed symptomatic endometriosis as further defined by the inclusion/exclusion criteria described in the study protocol.

***Intercurrent events (IEs)*:**

a) Early discontinuation of study intervention:
   - The IEs of premature discontinuation of study treatment due to a non-COVID-19-related AE and a premature discontinuation of study treatment due to a lack of efficacy will be considered as treatment failure and as such addressed by the composite strategy. The baseline value and the pre-discontinuation value will be considered as representative of the unfavorable outcomes associated with these events, respectively. Returning to baseline pain level is considered to be a clinically plausible outcome for participants not being able or willing to take the study treatment because of an AE and is considered to be an unfavorable outcome. The pre-discontinuation value reflects the pain level that the participant feels unfavorable enough to discontinue the study treatment due to a lack of efficacy.

     For premature discontinuation of study treatment due to a non-COVID-19-related AE, the baseline value will be imputed as:

a) If at least 15 observations are collected in the cycle in which the IE occurs: a weighted average of the observed values and the baseline value will be calculated for the current cycle, with the weights being number of available days divided by 28 days and number of missing days divided by 28 days. The baseline value will be used for the subsequent cycles.

b) If less than 15 observations are collected in the cycle in which the IE occurs: the baseline value will be used for this and the subsequent cycles.

For premature discontinuation of study treatment due to a lack of efficacy the last observation will be carried forward.

- The IEs of early discontinuation of the study intervention due to any other reason, including COVID-19 related reasons, will be addressed by the hypothetical strategy. The assumed hypothetical scenario is that participants would not discontinue the study treatment at that time and would have similar outcomes as other ptextarticipants in the study. If a participant continues completing the e-diary, the information included after the IE will be discarded.

b) The IEs of non-compliance with study intervention will be addressed using the treatment policy strategy: compliance will be calculated for each period between two consecutive visits. If the compliance is too low (i.e. below 80%), or too high (i.e. >120%, which is highly unlikely) for some but not all post-baseline cycles, the data collected during these no-compliant periods will be used as observed.

c) Increased and decreased intake of standardized rescue medication taken for EAPP and taken for any other reason than EAPP will be addressed by the treatment policy strategy. Earlier trials suggest that the intake of standardized rescue medication taken for EAPP is expected to be low. Any change in intake of standardized rescue medication taken for either EAPP or other reasons is considered as part of a general background treatment. Consequently, the treatment policy strategy will be used, and the standardized rescue medication is specified as part of the treatment attribute.

Other potential IEs (e. g. pregnancy or death due to any cause, including COVID-19 infection, or surgery) are considered to be unlikely. The occurrence of unexpected IEs will be evaluated during the blinded review and the estimand definition will be amended if needed.

6.2.1.1 Primary efficacy Analysis

The primary analysis will be performed on the pPPS analysis set as defined in Section 5.1.

For the primary efficacy variable, it is planned to perform a test for a dose-response signal in the study population, under the assumption of a monotone dose-response relationship in the dose range considered. The MCP-Mod method [5] combining multiple comparison *p*rocedures principles with *mod*eling techniques (MCP-Mod) will be used for the primary analysis of the primary efficacy variable. This method allows the flexibility of modeling for dose estimation, while preserving the robustness to model misspecification associated with MCP procedures. More specifically, it is planned to use a generalization of the original MCP-

Mod method which allows to perform dose-response testing and modeling in conjunction with the response variable being described by a parametric model [6].The key idea of this generalization is to decouple the dose-response model from the expected response so that the dose response can be characterized using a suitable parameter in the probability distribution of the response.

Assumptions

It is assumed that the primary efficacy variable, denoted as Y, is measured at the patient level for the 4 parallel groups corresponding to dose levels: (placebo =) $d1 < d2 < \ldots < d$K, where K = 4.

Furthermore, assume that the probability model for the patient responses Y includes parameters $\mu_{d_1}, \ldots, \mu_{d_K}$ for the changes from baseline in the primary endpoint capturing the dose-response effect for the doses $d1$, $d2$, … , dK. This probability model may depend on other (nuisance) parameters and covariates and is described in detail further below.

Subsequently, it is assumed that the dose-response parameters μ_{d_k}, $k = 1, \ldots, K$ are related through a dose-response model $\mu_d = f(\mathrm{d}, \boldsymbol{\theta})$ where $f(.)$ is parameterized by a vector of parameters $\boldsymbol{\theta}$.

Let $\widehat{\boldsymbol{\mu}} = (\hat{\mu}_{d_1}, \ldots, \hat{\mu}_{d_K})$ denote the vector of estimated dose-response parameters obtained from a parametric model of Y. The key assumption for the generalized MCP-Mod method is that $\widehat{\boldsymbol{\mu}}$ has an approximate distribution $N(\boldsymbol{\mu}, \boldsymbol{S})$, where $\boldsymbol{S}$ denotes the variance-covariance of $\widehat{\boldsymbol{\mu}}$. The estimates for $\widehat{\boldsymbol{\mu}}$ and $\widehat{\boldsymbol{S}}$ are obtained first, while the estimation of $\boldsymbol{\theta}$ is done in a separate, second stage based on $\widehat{\boldsymbol{\mu}}$ and $\widehat{\boldsymbol{S}}$.

The candidate set of models consists of the $M = 4$ models given by

$$f(d, \boldsymbol{\theta}) = \theta_0 + \theta_1 f^0(d, \boldsymbol{\theta}^0) = -2.3 + E_{max} d^{\eta} / \left(ED_{50}^{\eta} + d^{\eta}\right),$$

where d is the dose ranging from 0 mg to 150 mg, θ_0 is a location parameter, and θ_1 is a scale parameter such that only $\boldsymbol{\theta}^0$ determines the shape of the model function, where $\boldsymbol{\theta}^0$ is used to denote parameters associated with $f^0(.)$ component. Guestimates for parameters $\boldsymbol{\theta}$ assumed at the trial design stage are listed in Table 1. The two E_{max} models and the two sigmoidal E_{max} models assume a strictly monotonically decreasing dose-response, the illustration of which is provided in the Figure 9–1 in the clinical study protocol.

Table 1: Parameters of the dose-response-curves in the candidate set

Model		Parameters		
Abbreviation	Indicator m	ED_{50}	E_{max}	Hill factor η
emax1	1	25	-1.4	1
emax2	2	100	-2.0	1
sigEmax1	3	35	-1.2	3
sigEmax2	4	75	-1.2	5

Analysis

Step 1: Estimation of $\boldsymbol{\mu}$ and $\boldsymbol{S}$ from a mixed model for repeated measures with covariate adjustment

The expected dose response at the tested dose levels will be captured by the adjusted model-based treatment group means of the primary variable's values obtained from a mixed model for repeated measures (MMRM). Absolute change in mean worst EAPP from baseline to each post-baseline time point will be modeled as the dependent variable. The model will include fixed effects for treatment group (four treatment groups: placebo group and 25 mg, 75 mg, and 150 mg BAY 1817080), time point (Visits 3, 4, and 5), time point by treatment group interaction, region (Japan vs. ROW excluding China), and baseline EAPP as continuous covariate. An unstructured variance-covariance pattern will be used to model variance-covariance of the within-subject errors. This variance-covariance matrix will be estimated across treatment groups. In case the model does not converge, alternative variance-covariance structures will be considered. Variance-covariance parameters will be estimated using Restricted Maximum Likelihood (REML) with the Newton-Raphson algorithm and using Kenward-Roger method for calculating the denominator degrees of freedom and adjusting standard errors.

Least squares means for each treatment group at Visit 5 will be estimated from the time point by treatment group interaction term together with their variance-covariance matrix and will represent $\widehat{\boldsymbol{\mu}}$ and $\widehat{\boldsymbol{S}}$, respectively. The 80% CIs will also be estimated.

Absolute change in mean worst EAPP from baseline to post-baseline Visits 3, 4, and 5 may be missing for some participants due to a lack of compliance with eDiary completion. All such missing data and data that are unobservable after IEs intended to be addressed by a hypothetical strategy, as described in the primary estimand definition in Section 6.2.1, will be assumed to be Missing at Random (MAR) and will be modeled as such using the MMRM described above. That is, values of subjects with missing and unobservable data will be assumed to be distributed similarly to subjects with observed data in their treatment group conditional on the covariates included in the model and on partially observed data.

MMRM analysis will be implemented using SAS Proc MIXED.

Step 2: Detection of dose-response signal

For detecting an overall trend, or a dose-response signal, each of the $M = 4$ dose-response shapes in the candidate set will be tested, using a single contrast test.

For each model *m* in the candidate set

the null hypothesis $H_{0m}: (\boldsymbol{c}_m)'\boldsymbol{\mu} = 0$

will be tested against

the respective 1-sided alternative hypothesis, $H_{1m}: (\boldsymbol{c}_m)'\boldsymbol{\mu} > 0$

where $\boldsymbol{c}_m = (c_{m1}, \ldots, c_{mK})'$ is the optimal contrast vector representing model *m*.

The optimum contrast coefficients and critical values for the four contrast tests on the dose-response shapes will be derived based on the guestimates for parameters $\boldsymbol{\theta}^0$ of standardized versions of the models in the candidate set specified at the design stage, the actual sample size for each treatment group after study completion, and the covariance matrix $\widehat{\boldsymbol{S}}$ estimated from actual data using the MMRM described above [6].

Each dose-response model *m* will be tested using a single contrast test:

$$T_m = (\boldsymbol{c}_m)' \, \hat{\boldsymbol{\mu}} / \left[(\boldsymbol{c}_m)' \hat{\boldsymbol{S}} \boldsymbol{c}_m \right]_{\boldsymbol{m,m}}^{\mathbf{1/2}}$$

where $[\boldsymbol{A}]_{\boldsymbol{m,m}}$ denotes the m[th] diagonal element of the matrix $\boldsymbol{A}$.

The final detection of a dose-response signal is based on the maximum contrast test statistic $T_{max} = \max(T_1, \ldots, T_4)$ and can be concluded if $T_{max} > q_{1-\alpha}$, where $q_{1-\alpha}$ is the multiplicity adjusted critical value at level $\alpha = 0.1$. The MCP-Mod method takes multiplicity into account, and no further multiplicity adjustments are needed.

If no candidate model is statistically significant, the procedure stops indicating that a dose-response relationship cannot be established from the observed data. If, on the other hand, there is one or more models with a significant adjusted *p* value, these significant models will be retained and fitted as described in the next step.

Step 3: Modeling and estimation of target doses

The selected dose-response models (with a significant dose-response trend detected at the previous step) will be fitted to $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{S}}$ to estimate the model parameters $\boldsymbol{\theta}$ using a generalized least squares (GLS) estimation criterion [6].

If multiple models have significant adjusted p-values at Step 2 and can be successfully fitted (without convergence or numerical stability issues), final inference will be based on the model with the best fit as determined by the AIC criterion.

To estimate the target dose(s) of interest the minimal effective dose (MED) will be determined:

$$MED = argmin_{(\mathrm{d=d1,d4})} \{ \tilde{\mathrm{f}}(d, \hat{\theta}) > \tilde{\mathrm{f}}(d_1, \hat{\theta}) + \Delta \}$$

where $\tilde{\mathrm{f}}(d, \hat{\theta})$ is the best estimated model and Δ is the smallest relevant difference, by which a dose is expected to be better than placebo in terms of efficacy measured by the primary efficacy variable.

The MED will be provided for the following smallest relevant differences:

Δ = clinically meaningful threshold (e.g. -1)

Δ = the observed effect of Elagolix in the study

The final choice of the target dose will be based on the analyses of the primary, secondary, and other efficacy variables, as well as safety considerations

MCP-Mod related steps 2 and 3 will be performed using the R "DoseFinding" package (version 0.9-17 or higher) [7].

The level of significance for statistical testing is set to 10 % one-sided. The type I error will be controlled for the primary analysis only.

In consistency with the one-sided level of significance given the generally two-sided reporting of CIs, 80 % CIs will be reported for the primary analysis and its related sensitivity/supplementary analyses in addition to 95 % CIs.

6.2.2 Secondary efficacy endpoint

The same estimand as the primary efficacy estimand - except the endpoint - will be used for the secondary efficacy endpoints analyses. The IEs a), b) and c) will be handled as described in section 6.2.1.

For the identification of at least 1 superior effective dose of BAY 1817080 compared to placebo in terms of the absolute change in mean worst EAPP from baseline to end of intervention, estimates and CIs will be derived from MMRM. The mean worst EAPP at baseline, the region (Japan vs. ROW excluding China) visit and treatment*visits are set as covariates.

Summaries will be provided by treatment group and overall. Analyses will be performed including the Elagolix population arm.

6.2.3 Other endpoints

The descriptive analyses of the other endpoints described in the following subsections will be based on the FAS and presented by treatment group.

6.2.3.1 Endpoints on the worst EAPP

Absolute change in mean worst EAPP from baseline to the first 4 weeks/8 weeks of intervention/end of intervention (measured daily on the NRS by item 1 of the ESD)

Absolute change in mean worst EAPP from end of intervention to non-overlapping, consecutive 28-day intervals during follow-up (measured daily on the NRS by item 1 of the ESD)

Absolute change in mean worst EAPP on bleeding days from baseline to the first 4 weeks/8 weeks of intervention/end of intervention (measured on the NRS by item 1 and item 4 of the ESD)

Absolute change in mean worst EAPP on bleeding days from end of intervention to non-overlapping, consecutive 28-day intervals during follow-up (measured on the NRS by item 1 and item 4 of the ESD)

Absolute change in mean worst EAPP on non-bleeding days from baseline to the first 4 weeks/8 weeks of intervention/end of intervention (measured on the NRS by item 1 and item 4 of the ESD)

Absolute change in mean worst EAPP on non-bleeding days from end of intervention to non-overlapping, consecutive 28-day intervals during follow-up (measured on the NRS by item 1 and item 4 of the ESD)

6.2.3.2 Rescue medication

The overall amount of rescue medication during the days considered for pain evaluation will be summarized by treatment group and by 28-day time window.

6.2.3.3 Other HRQoL and PRO associated endpoints

- Descriptive assessment will be done on the basis of the VAS and on dimension level assessment basis.

  Frequency tables providing the number and percentages of the different response options to individual questions will be displayed. In addition, changes from baseline for the single questions will be displayed with shift tables by treatment.

  The EQ visual analogue scale (VAS) values of the EQ-5D-5L and their changes to ue baseline will be summarized by treatment as continuous variable; arithmetic mean, SD, minimum, quartiles, median and maximum will be displayed by treatment.

  The scoring algorithm described in the EQ-5D-5L manual will be used.
- EIS single items, EIS physical subscale, EIS emotional subscale, and the eight remaining EIS individual items during baseline, at weeks 4/8/end of

intervention/FUP1/FUP2 and absolute change from each week after start of intervention to baseline. Full details of scoring algorithms and transformation of domain scores and single item scores are provided in Section 8 Appendix 1.

- VAS for pelvic pain at baseline, at weeks 4/8/end of intervention/FUP1/FUP2 and absolute change from each visit after start of intervention to baseline
- Patient Global Impression of Change (PGI-C) and change from baseline in Patient Global Impression of Severity (PGI-S), at visit 3, 4 and 5. Frequency tables providing the number and percentages of the different response options to individual questions will be displayed for PGI-S and PGI-C. In addition, changes from baseline for PGI-S and changes from Visit 3 for PGI-C will be displayed using shift tables for the single questions.
- PAQ (Productivity Activity Questionnaire). single items at screening and end of intervention; absolute change of PAQ individual items from end of intervention to screening
- PainDETECT single items and total score at baseline, end of intervention/FUP1/FUP2 and absolute change from end of intervention/end of follow-up to baseline for numeric items and shift tables for categorical items.
- PCS single items, subscales, and total score at baseline
- Absolute change in PSQ3 single items from end of intervention to baseline

These summary statistics and difference to baseline will be presented by treatment.

6.3 Pharmacokinetics/pharmacodynamics (PK/ PD)

BAY 1817080 plasma concentrations will be summarized by intervention group, visit, and planned sampling time. Box-plots showing plasma concentration vs timepoint will be created for each intervention, highlighting patients who fullfil the criteria for close liver observation.

6.4 Safety

All variables in this section will be analyzed descriptively based on the SAF.

6.4.1 Adverse events (AEs)

All AEs will be collected from signing the informed consent until the SFU visit.

A treatment-emergent AE (TEAE) is defined as any event arising or worsening after the start of study drug administration until 14 days after the last study medication intake.

The incidence of TEAEs will be assessed as secondary safety endpoint. The incidence of AEs and TEAEs will be analyzed by descriptive statistics, such as frequency tables.

An overall summary of the number and percentage of participants with AEs and TEAE will be presented. In addition, all AEs and TEAEs will be tabulated according to the affected system organ class and preferred term, as coded by the Medical Dictionary for Regulatory Affairs (MedDRA).

The frequency of the AEs related to potential and identified risks of Eliapixant will be displayed. Regarding the identification specifications for these AEs, the following SMQs and PTs will be applied:

1. SMQ “Taste and smell disorders”

2. SMQ "Hemorrhages"
3. MLG Hypotension
    - PT: Blood pressure ambulatory decreased;
    - PT: Blood pressure decreased;
    - PT: Blood pressure diastolic decreased
    - PT: Blood pressure systolic decreased;
    - PT: Diastolic hypotension
    - PT: Hypotension;
    - PT: Mean arterial pressure decreased
    - PT: Heart rate decreased
    - PT: Bradyarrhythmia
    - PT: Bradycardia
    - PT: Sinus bradycardia
    - PT: Dizziness
    - PT: Dizziness postural
    - PT: Dizziness exertional
    - PT: Syncope
    - PT: Presyncope
    - PT: Loss of consciousness
    - PT: Blood pressure orthostatic decreased;
    - PT: Orthostatic hypotension;
4. SMQ "Drug related hepatic disorders - comprehensive search"

Further tables will be provided for serious and/or drug related TEAEs. Tables for non-serious TEAEs will also be provided.

In addition, a separate table summarizing TEAEs that occurred in more than 5% of the subjects will be provided.

The incidence of all AEs during pre-treatment and during post-treatment (that is, AEs occurring more than 14 days after end of treatment with study drug) will be tabulated separately.

The summaries will be tabulated by intervention group and overall.

Serious adverse events, deaths and adverse events leading to discontinuation will be listed. The date, relative day (to study intervention) and treatment-emergent flag will be included.

Further summaries of adverse events by intensity and worst outcome, may be provided, consistent with Bayer Global Medical Standards.

6.4.1.1 Taste-related and smell-related AEs

Impairment or loss of smell and/or taste is a common symptom of COVID-19. Taste-related/ smell-related AEs as well as the responses to the questions of the taste/smell disturbance assessments will be analyzed using descriptive statistics.

To distinguish P2X3 related taste AEs from COVID-19 related smell/taste AEs the details of smell/taste disturbance assessments will be evaluated together with the SARS-CoV-2/serology IgG test results, the AEs confirming COVID-19 and the timing information by individual patient review.

SMQ "Taste and smell AEs" has 14 MedDRA PTs, these will be devided in two cateogires one for "Taste AEs" and one for "Smell-Related AEs" as follows:

The following preferred terms will be combined into the "Taste AE" - category:

PT: Ageusia

PT: Dysgeusia

PT: Gustometry abnormal

PT: Hallucination, gustatory

PT: Hypergeusia

PT: Hypogeusia

PT: Taste disorder

The following preferred terms will instead be included in the "Smell-related AEs" category:

PT: Anosmia

PT: Congenital anosmia

PT: Hallucination, olfactory

PT: Hyposmia

PT: Olfactory nerve disorder

PT: Olfactory test abnormal

PT: Parosmia

Taste- and smell-related AEs, along with responses about frequency and bothersomeness, will be analyzed using descriptive statistics.

In addition, the dose-dependency of taste-related AEs will be assessed using a logistic regression model with treatment as factor. The odds ratio and its 80% and 95% confidence intervals, and the p-value for treatment comparisons from the logistic regression model will be provided. Only taste-related AEs that cannot be attributed to a SARS-CoV-2 infection will be considered for the model.

6.4.2 Study Treatment Duration and Exposure

Study treatment duration and the average daily dose will be summarized overall for SAF. The duration of study treatment (in days) is derived by the following formula:

last dose date – first dose date + 1- number of days treatment was interrupted due to COVID-19 pandemic-related reasons.

Descriptive statistics of treatment duration will be presented.

The number of tablets taken will be summarized descriptively by treatment group and overall.

Using the drug accountability data, the average daily dose for overall study is calculated for subjects on active treatment based on the following formulas depending on the treatment arm:

For the 25mg treatment arm

$$25\text{mg} \cdot \frac{\frac{1}{3} \cdot \text{Total Number of 25mg tablets taken}}{\min(\text{last tablets returned date},\ \text{last treatment intake date}) - \text{first tablets dispensed date}}$$

For the 75mg treatment arm

$$25\text{mg} \cdot \frac{\text{Total Number of 25mg tablets taken}}{\min(\text{last tablets returned date},\ \text{last treatment intake date}) - \text{first tablets dispensed date}}$$

For the 150mg treatment arm

$$25\text{mg} \cdot \frac{\frac{2}{3} \cdot \text{Total Number of 25mg tablets taken}}{\min(\text{last tablets returned date},\ \text{last treatment intake date}) - \text{first tablets dispensed date}}$$

$$+100\text{mg} \cdot \frac{\text{Total Number of 100mg tablets taken}}{\min(\text{last tablets returned date},\ \text{last treatment intake date}) - \text{first tablets dispensed date}}$$

For the Elagolix arm

$$150\text{m} \cdot \frac{\text{Total Number of 150mg Elagolix tablets taken}}{[\min(\text{last tablets returned date},\ \text{last treatment intake date}) - \text{first tablets dispensed date}) + 1]}$$

Overall Compliance will be calculated as

$$\frac{\text{Total Number of 25mg tablets taken} + 4 \cdot \text{Total Number of 100mg tablets taken}}{14 \cdot (\min(\text{last tablets returned date},\ \text{last treatment intake date}) - \text{first tablets dispensed date})}.$$

In addition, overall Compliance for the Elagolix arm will be calculated as

$$\frac{\text{Total number of 150mg Elagolix tablets taken}}{[\min(\text{last tablets returned date},\ \text{last treatment intake date}) - \text{first tablets dispensed date}) + 1]}.$$

Compliance between visits will be calculated as

$$\frac{\text{Number of 25mg tablets taken between visits} + 4 \cdot \text{Number of 100mg tablets taken between visits.}}{14 \cdot (\min(\text{tablets returned date; last treatment intake}) \text{ at Visit x} - \text{ tablets dispensed date at Visit } (x - 1))}$$

If compliance between two consecutive visits cannot be determined (e.g., if a participant does not return medication at the scheduled time), then compliance will be averaged across the periods that are affected.

For between-visit compliance, periods from unscheduled visits are assigned to nearest scheduled between visit periods.

The compliance will be summarized descriptively by treatment group and overall. In addition, compliance will be categorized into three groups (< 80%, 80-120%, > 120%) and summarized by treatment group and overall.

6.4.3 Liver Function Laboratory Parameters

The number of participants with laboratory abnormalities will be provided for SAF for the following liver function laboratory parameters. The table will be provided for both baseline and post-baseline measurements. The measurements from both scheduled / unscheduled visits will be used. The ratio to the upper limits of normal (ULN) will be used for these parameters.

- Alanine aminotransferase (ALT): > 2xULN – ≤3xULN, > 3xULN – ≤ 5xULN, > 5xULN – ≤ 8xULN, > 8xULN
- Aspartate aminotransferase (AST): > 2xULN – ≤3xULN, > 3xULN – ≤ 5xULN, > 5xULN – ≤ 8xULN, > 8xULN
- Total bilirubin (TBL): > 2xULN
- alkaline phosphatase (AP): ≥ 1.5xULN
- γ-GT (gamma GT): ≥ 2xULN
- International normalized ratio (INR): > 1.5xULN

In addition, frequency tables based on categories above and box plot for measurements over time by visit will be provided.

The number of participants with postbaseline elevated liver values according to the following criteria will be provided:

- > 3xULN of ALT or AST accompanied (using the values taken from the same serum sample) by > 2xULN of TBL, or
- > 3xULN of ALT or AST accompanied by > 1.5xULN of INR, or
- participants who reported unspecific symptom (fatigue, nausea, vomiting, right upper abdominal quadrant pain or tenderness, fever, rash and/or eosinophilia) with the CRF pages: Clinical Signs and Symptoms with elevated liver enzymes / Clinical Signs and Symptoms with elevated liver enzymes (Follow-up)
    - For participants with eosinophilia, only the case if their measurement of laboratory parameter of Eosinophils/Leukocytes > 5% will be included into the analysis.

A subject listing will be provided for those participants.

The eDISH plot (maximum of ALT postbaseline value × maximum of TBL postbaseline value) will be provided. A line plot of AP, ALT, AST and TBL over time will be provided for participants with liver injury (ALT>5xULN).

6.4.4 Other safety variables

6.4.4.1 Laboratory parameters

Quantitative data (e.g. laboratory parameters) will be described by the following summary statistics: arithmetic mean, standard deviation, quartiles, median, minimum, and maximum. Safety parameters with categorical data will be summarized by reporting number and percent of participants under such categories.

These summary statistics and the difference to baseline (i.e. pre-dose measurements, performed before the first administration of the study medication) will be presented by intervention.

Subjects with abnormal laboratory values (values out of the reference range) will be summarized using shift tables comparing the baseline to post-baseline measurements by timepoint and also for the worst post-baseline measurement. The worst post baseline measurement includes all post baseline values including laboratory values from unscheduled visits.

Frequency tables will be provided for qualitative data. Laboratory data outside the reference range will be listed and flagged with 'L' for low and 'H' for high. Additional tables with all abnormal values will be presented.

The results of the SARS-CoV-2 tests will be displayed descriptively.

6.4.4.2 Menstrual bleeding pattern

For the analysis of bleeding pattern, the following measures will be derived based on ESD item 4 and analyzed using descriptive statistical methods overall and by treatment group:

1) number of bleeding / spotting days

Number of days where ESD item 4 is spotting, light, normal, or heavy within the respective 28-day reference period.

2) number of bleeding days (excluding spotting-only days)

Number of days where ESD item 4 is light, normal, or heavy within the respective 28-day window.

3) number of spotting-only days

Number of days where ESD item 4 is spotting within the respective 28-day window.

4) number, mean length, and maximal length of bleeding / spotting episodes

According to (2), an episode is defined as consecutive days where ESD item 4 is spotting, light, normal, or heavy, preceded and followed by at least 2 days without bleeding, i.e. ESD item 4 is none. The number, mean length and maximal length for the respective 28-day window will be calculated.

5) number, mean length, and maximal length of spotting-only episodes

According to (2), an episode is defined as consecutive days where ESD item 4 is spotting, preceded, and followed by at least 2 days without bleeding, i.e. ESD item 4 is none. Episodes of length 1 are possible. The number, mean length and maximal length for the respective 28-day window will be calculated.

6.4.4.3 Cervical smear

A frequency table for cervical smear findings will be produced by treatment group and visit (Visit 1 and FUP). A listing will be provided with the cytology Cervical smear High risk HPV-DNA results at screening visit and at FUP visit for each participant.

6.4.4.4 Pregnancy testing

Both Highly sensitive urine pregnancy test (β-hCG) and urine pregnancy tests will be summarized by the number of women taking a pregnancy tests and the test results.

A listing of pregnancy tests which summarizes the date of pregnancy test, the type (i. e., urine, serum, or home pregnancy test) and the result will be summarized. Listing of participants with positive pregnancy tests will be provided.

6.4.4.5 Vital signs

Vital signs (heart rate, systolic blood pressure, diastolic blood pressure and weight) will be summarized by treatment group and visit grouped by treatment period, including absolute change from baseline where appropriate. Boxplots for absolute change from baseline at each visit will be presented by treatment group.

7. Document history and changes in the planned statistical analysis

SAP V 1.0 dated 31 MAR 2021.

Approval of the SAP V 1.0 dated 08 APR 2021.

SAP final version 2.0 dated 23 MAY 2022. SAP was updated to take into consideration project and study early termination. For this reason, study objectives were corrected. Interim analyses section was modified to reflect the early termination of the study.

Handling and Missing data sections was updated to reflect missing end treatment date in eCRF. Data Rules section was updated to reflect ePRO questionnaires being pushed to participants not in line with start of treatment date. Data rules for laboratory parameters were also modified to reflect observing values higher and lower than the preselected limits.

In the primary efficacy analysis, the estimand section was expanded to provide more clarification on how to impute values after the occurrence of an intercurrent event. Subgroup analyses and sensitivity analyses were removed from the SAP. The actigraphy associated endpoint analyses were removed from the SAP.

SMQs for the specification of AEs were modified to be aligned on compound level.

SAP was also updated to correct categories for liver function laboratory parameters and add analyses of close liver observations. In addition, the list of preferred MedDRA terms relevant to COVID-19 has been updated. The compliance calculation has been adapted to cover specific cases. In addition, the compliance definition for overall compliance and between-visit compliance was updated to be aligned on compound level.

Finally, an appendix section containing the derivation and the scoring of the EIS questionnaire was also included.

SAP V 2.0 dated 23 MAY 2022.

Approval of the SAP V 2.0 dated 23 MAY 2022.

SAP final version 3.0 dated 01 JULY 2022. SAP was updated to correct typos. All corrections included in this updated version of the SAP did not affect the primary and secondary analyses.

8. Supporting Documentation

8.1 Appendix 1: Scoring and Transformation of EIS

The following sections are taken from the EIS User Manual Version 3_0 [6].

8.1.1 Summary of EIS scoring algorithms

- EIS Physical activities domain score (based on EIS items 1-7)
- EIS Emotional well-being domain score (based on EIS items 8-14)
- EIS single-item scores, including:
    - Impact on sexual activities (EIS item 15)

- Limited enjoyment of sexual intercourse (EIS item 16)
- Guilt about avoidance of sexual intercourse (EIS item 17)
- Difficulty concentrating (EIS item 18)
- Difficulty sleeping (EIS item 19)
- Impact on household activities (EIS item 20)
- Impact on paid work or study (EIS item 21)
- Impact on social and leisure activities (EIS item 22)

All EIS domain and individual item scores range from 0 (indicating no impact) and 100 (indicating maximum impact). A summary of EIS scoring algorithms is provided in the table below.

Score	Item(s) employed	Calculation of score
EIS Physical Activity score	EIS items 1-7	When all 7 items are completed and item 7 is applicable (i.e., > 0): (100x((SUM EIS 1, 2, 3, 4, 5, 6, 7)-7)/28) When all 7 items are completed but item 7 is not applicable (100x((SUM EIS 1, 2, 3, 4, 5, 6)-6)/24); summed over the 4-week window, divided by the number of non-missing weeks within that 4-week window
EIS Emotional Well-being score	EIS items 8-14	(100x ((Sum EIS 8, 9, 10, 11, 12, 13, 14)-7)/28) summed over the 4-week window, divided by the number of non-missing weeks within that 4-week window
4 week mean of impact on sexual activities	EIS item 15	When item 15 is applicable (i.e., > 0): (100x (EIS item 15 -1)/4) summed over the 4-week window, divided by the number of non-missing (or applicable) weeks within that 4-week window.
4 week mean of limited enjoyment of sexual intercourse	EIS item 16	When item 16 is applicable (i.e., > 0): (100x (EIS item 16 -1)/4), summed over the 4-week window, divided by the number of non-missing (or applicable) weeks within that 4-week window.
4 week mean of guilt about avoidance of sexual intercourse	EIS item 17	When item 17 is applicable (i.e., > 0): (100x (EIS item 17-1)/4), summed over the 4-week window, divided by the number of non-missing (or applicable) weeks within that 4-week window.
4 week mean of difficulty concentrating	EIS item 18	(100x (EIS item 18 -1)/4), summed over the 4-week window, divided by the number of non-missing weeks within that 4-week window.
4 week mean of difficulty sleeping	EIS item 19	(100x (EIS item 19 -1/4), summed over the 4-week window, divided by the number of non-missing weeks within that 4-week window.
4 week mean of impact on household activities	EIS item 20	(100x (EIS item 20 -1)/4), summed over the 4-week window, divided by the number of non-missing weeks within that 4-week window.
4 week mean of impact on paid work or study	EIS item 21	When item 21 is applicable (i.e., > 0): (100x (EIS item 21 -1)/4), summed over the 4-week window, divided by the number of non-missing (or applicable) weeks within that 4-week window.

Score	Item(s) employed	Calculation of score
4 week mean of impact on social or leisure activities	EIS item 22	(100x (EIS item 22 -1)/4), summed over the 4-week window, divided by the number of non-missing weeks within that 4-week window

8.1.2 Transformation of scores

The sum of all responses given to a set of items is referred to as a score. For the EIS, scores are transformed to a 0-100 scale (where 0 represents no impact and 100 represents maximum impact) so to facilitate comparison across domains and items.
The general formula for transformation of multi-item domain scores (physical activities and emotional well-being) for each respondent is as follows:

> (100 x (sum of participant responses to all applicable items – number of applicable items)) / (maximum potential sum of responses to all applicable items – number of applicable items)

The general formula for transformation of single item scores (EIS items 15-22) for each respondent, where applicable (i.e., scores > 0) is as follows:

> (100 x (item response - 1)) / (maximum potential item response-1)

9. References

[1] R Core Team, R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, 2020.

[2] Best Practice Document RD-SOP-1107: Recording and Evaluation of Bleeding Data, version 10.0.

[3] RD-M-0001: COVID_CAT_05_SPA related tasks to COVID19-pandemic v6.0," 2021.

[4] Operational Instruction RD-OI-0117: Develop specifications for data collection, cleaning and review tools, version 2.0

[5] Bretz F, Pinheiro JC, Branson M. Combining multiple comparisons and modeling techniques in dose-response studies. Biometrics. 2005;61(3):738-48.

[6] Pinheiro JC, Bornkamp B, Glimm E, Bretz F. Model-based dose finding under model uncertainty using general parametric models. *Statistics in Medicine*. 2014;33:1646–1661.

[7] Choi et al. (2011): Choi L, Liu Z, Matthews CE, Buchowski MS. Validation of accelerometer wear and nonwear time classification algorithm. Med Sci Sports Exerc. 2011 Feb;43(2):357-64.

[8] Tracy et al. (2014): Tracy DJ, Xu Z, Choi L, Acra S, Chen KY, et al. (2014) Separating Bedtime Rest from Activity Using Waist or Wrist-Worn Accelerometers in Youth. PLoS ONE 9: e92512.

[9] Tracy et al. (2018): Tracy JD, Acra S, Chen KY, Buchowski MS. Identifying bedrest using 24-h waist or wrist accelerometry in adults. PLoS One. 2018 Mar 23;13(3):e0194461.

[10] (Fekedulegn et al., 2020): Fekedulegn D, Andrew ME, Shi M, Violanti JM, Knox S, Innes KE. Actigraphy-Based Assessment of Sleep Parameters, *Annals of Work Exposures and Health*, Volume 64, Issue 4, May 2020, Pages 350–367

[11] Cole et al. (1992): Cole R, Kripke D, Gruen W, Mullaney D, Gillin J (1992) Automatic sleep/wake identification from wrist activity. Sleep 15: 461±469.

[12] de Souza et al. (2003): de Souza L, Benedito-Silva AA, Pires ML, Poyares D, Tufik S, Calil HM. Further validation of actigraphy for sleep studies. Sleep. 2003 Feb 1;26(1):81-5.

[13] Staudenmayer et al. (2015): Staudenmayer J, He S, Hickey A, Sasaki J, Freedson P. Methods to estimate aspects of physical activity and sedentary behavior from high-frequency wrist accelerometer measurements. J Appl Physiol (1985). 2015 Aug 15;119(4):396-403.

[14] (ActiGraph, 2016): "Moving Average Vector Magnitude (v.1) Step Algorithm." ActiGraph white paper. ActiGraph, LLC. September 2016.

[15] (Actigraph, 2019): "University of West Florida Step Counting Algorithm." ActiGraph white paper. ActiGraph, LLC. August 2019.