STATISTICAL ANALYSIS PLAN

Study Title:	A Phase 2a/2b Randomized Double-Blind Placebo-Controlled Study to Evaluate the Efficacy and Safety of V**o**lixibat in Adult Women with Intra**h**epatic Cholestasis of Pregnancy and Elevated Serum Bile **A**cid Co**n**centr**a**tions (OHANA)
Name of Test Drug:	Volixibat
Study Number:	VLX-401
Protocol Version (Date):	Original Protocol, Version 1, 14 October 2020
Analysis Type:	Interim and Final Analyses
Analysis Plan Version:	1.0
Analysis Plan Date:	13 November 2020
Analysis Plan Author(s):	Will Garner, PhD

CONFIDENTIAL AND PROPRIETARY INFORMATION

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ABBREVIATIONS

7αC4	7α-hydroxy-cholesten-3-one
AE	adverse event
AESI	adverse event of special interest
ALP	alkaline phosphatase
ALT	alanine aminotransferase
ANCOVA	analysis of covariance
AST	aspartate aminotransferase
ATC	Anatomical Therapeutic Chemical
AUC	area under the plasma concentration-time curve
BMI	body mass index
BPP	biophysical profile
CMH	Cochran-Mantel-Haenszel
CI	confidence interval
C_{max}	maximum observed concentration
CSR	clinical study report
CSS	clinician scratch scale
CTCAE	Common Terminology Criteria for Adverse Events
CTG	cardiotocography
EAIR	exposure-adjusted incidence rate
ECG	electrocardiogram
eCRF	electronic case report form
ECI	events of clinical interest
eDiary	electronic diary
EQ-5D-5L	EuroQoL 5-dimension, 5-level
FGF-19	fibroblast growth factor 19
GLP-1	glucagon-like peptide 1
HLGT	high-level group term
HLT	high-level term
ICF	informed consent form
ICP	intrahepatic cholestasis of pregnancy
ID	identification
IDMC	independent data monitoring committee
INR	international normalized ratio
ItchRO	Itch Reported Outcome
ITT	intent to treat
IV	intravenous
IXRS	interactive voice or web response system
LLT	lower-level term
LMP	last menstrual period

LOQ	limit of quantitation
LR	logistic regression
LS	least squares
MAR	missing at random
MedDRA	Medical Dictionary for Regulatory Activities
MH	Mantel-Haenszel
MI	multiple imputation
mITT	modified intent-to-treat
MMRM	mixed-effects model for repeated measures
MNAR	missing not at random
NICU	neonatal intensive care unit
NST	nonstress test
PD	pharmacodynamic
PGIC-Itch	Patient Global Impression of Change-Itch
PIS-Itch	Patient Impression of Severity-Itch
PK	pharmacokinetic
PP	per-protocol
PPROM	preterm premature rupture of the membranes
PRO	patient-reported outcome
PROMIS	Patient-Reported Outcomes Measurement Information System
PT	preferred term
Q1, Q3	first quartile, third quartile
REML	restricted maximum likelihood
SAE	serious adverse event
SAP	statistical analysis plan
sBA	serum bile acid
SD	standard deviation
SF	short form
SOC	system organ class
TEAE	treatment-emergent adverse event
TFLs	tables, figures, and listings
T_{max}	time to reach C_{max}
TRAE	treatment-related treatment-emergent adverse event
ULN	upper limit of normal
VLX	volixibat
WHO	World Health Organization

1. INTRODUCTION

This statistical analysis plan (SAP) describes the statistical analysis methods and data presentations to be used in tables, figures, and listings (TFLs) for the interim analysis as well as final analysis to support the clinical study report (CSR) for Study VLX-401. This SAP is based on the study protocol Version 1, 14 October 2020, and the electronic case report forms (eCRF). The SAP will be finalized before the interim analysis.

Any deviations from the approved protocol will be documented in either Section 6 (efficacy) and/or Section 7 (safety).

1.1. Study Objectives

The following objectives and endpoints will be evaluated in participants who meet study eligibility criteria. Unless otherwise specified, these apply to participants in the primary cohort.

1.1.1. Primary Objective and Endpoint

- To assess the efficacy of volixibat (VLX) on the reduction of elevated sBA concentrations in participants with intrahepatic cholestasis of pregnancy (ICP) on the basis of the following endpoint:
    - Mean change from baseline to Week 3 of the study treatment period in total serum bile acid (sBA) concentration

1.1.2. Secondary Objectives and Endpoints

- To assess the efficacy of VLX on pruritus due to ICP on the basis of the following endpoint:
    - Mean change from baseline to Week 3 of the study treatment period in the weekly average worst daily itch score as measured by the Adult Itch Reported Outcome (ItchRO)
- To assess the impact of volixibat on a composite perinatal outcome in participants with ICP on the basis of the following endpoint:
    - Proportion of participants experiencing one or more of the following:
        - Perinatal death, defined as in utero fetal death after randomization or known neonatal death up to 28 days after birth
        - Spontaneous preterm delivery, defined as delivery at <37 weeks' gestational age following spontaneous onset of preterm labor or preterm premature rupture of the membranes (PPROM)
        - Iatrogenic preterm delivery attributable to ICP or ICP-related complications, defined as delivery at <37 weeks' gestational age resulting from medical intervention (i.e., augmentation/induction of

labor or cesarean delivery without prior spontaneous onset of preterm labor or PPROM with one or more of the following as the indication for delivery):

- Diagnosis of ICP with or without suspected fetal compromise
- Persistently elevated/increasing sBA
- Worsening hepatic function
- Intolerable maternal ICP symptoms

– Neonatal unit admission for ≥12 hours from birth until hospital discharge for one or more of the following indications:

- Respiratory insufficiency/failure with requirement for oxygen supplementation
- Noninvasive or invasive respiratory support
- Hemodynamic instability or shock requiring clinical intervention
- Metabolic acidosis or signs of asphyxia
- Proven infection
- Hypoglycemia or feeding problems requiring intravenous (IV) fluids or tube feeding
- Meconium stained amniotic fluid and/or its sequelae (e.g., meconium aspiration, etc.)

1.1.3. Safety Objective and Endpoints

- To assess the safety and tolerability of VLX in participants with ICP on the basis of the following endpoints:
    - Incidence of treatment-emergent adverse events (TEAEs), serious adverse events (SAEs), adverse events of special interest (AESIs), events of clinical interest (ECIs), and AEs that lead to discontinuation of study drug
    - Incidence of clinically significant laboratory abnormalities

      Note: Safety and tolerability are primary objectives for the open-label supplemental cohort.

1.1.4. Exploratory Objectives and Endpoints

The following exploratory objectives and endpoints are applicable to both the primary and open-label supplemental cohorts:

- To assess volixibat drug levels in maternal and fetal plasma and pharmacodynamic (PD) markers in maternal blood in participants with ICP on the basis of the following endpoints:

  - Volixibat concentrations in maternal and fetal plasma (from umbilical cord sampling)

    Note: Assessment of maternal/fetal volixibat drug levels is a secondary objective for the open-label supplemental cohort.
  - Change in biomarkers of bile acid synthesis, inflammation, and pruritogens
  - Proportion of participants with sBA <40 μmol/L at end of the treatment period

- To assess the longer-term efficacy of volixibat in participants with ICP on the basis of the following endpoints:
  - Mean change from Week 3 to end of study treatment period in total sBA concentration
  - Mean change from baseline to end of study treatment period in total sBA concentration
  - Association between changes from baseline to Week 3 in total sBA and ItchRO as well as their association with the composite perinatal outcome
- To assess the effect of volixibat on additional perinatal outcomes on the basis of the following endpoints:
  - Mean gestational age at delivery
  - Proportion of participants with an early term delivery (37–38 weeks 6 days, inclusive)
  - Proportion of participants with a full-term delivery (≥39 weeks)
  - Proportion of participants experiencing a perinatal death
  - Proportion of participants with a spontaneous preterm delivery
  - Proportion of participants with an iatrogenic preterm delivery attributable to ICP
  - Proportion of neonates requiring neonatal intensive care unit (NICU) admission for ≥12 hours between birth and hospital discharge
  - Proportion of neonates with meconium-stained amniotic fluid at delivery
  - Proportion of participants requiring rescue medication for ICP
  - Proportion of neonates with Apgar score <7 at 5 minutes of life
  - Mean birth weight percentile
  - Mean placental weight percentile
  - Proportion of neonates with umbilical cord pH <7.0 at birth
  - Means for umbilical cord blood sBA, total cholesterol, LDL cholesterol, and glucose
  - Mean maternal estimated blood loss at delivery
  - Mean time from randomization to delivery

- To evaluate the effect of volixibat on additional measures of pruritus efficacy and quality of life in participants with ICP on the basis of the following endpoints:
    - Change from baseline in the 5-D Itch Scale
    - Change from baseline in Clinician Scratch Scale (CSS)
    - Change from baseline in EuroQoL 5-dimension, 5-level (EQ-5D-5L)
    - Change from baseline in Patient-Reported Outcomes Measurement Information System (PROMIS) Short Form (SF) Fatigue 7a
    - Change from baseline in PROMIS SF Sleep Disturbance
    - Change from baseline in Patient Impression of Severity-Itch (PIS-Itch)
    - Patient Global Impression of Change-Itch (PGIC-Itch)
- To assess the impact of volixibat on healthcare utilization in participants with ICP on the basis of the following endpoints:
    - Number and duration of medical care encounters, including surgeries and other selected procedures (inpatient and outpatient)
    - Dates and duration of hospitalization (total days or length of stay, including duration by wards [e.g., intensive care unit])
    - Number and type of diagnostic and therapeutic tests and procedures
    - Outpatient medical encounters and treatments (including physician or emergency department visits, tests and procedures, and medications)

1.2. Study Design

This is an operationally seamless adaptive clinical study that will consist of 2 parts.

Part 1: Proof of Concept, Dose Ranging, Pharmacokinetics, and Interim Analysis

After a screening period of up to 10 days, eligible participants will be identified for one of two cohorts, defined as follows:

- Primary cohort: Participants with ICP with a screening sBA level ≥20 μmol/L will be randomized with stratification (based on gestational age [<32 weeks or ≥32 weeks] at randomization, presence/absence of gestational diabetes, and highest sBA before randomization [<40 μmol/L, ≥40 μmol/L to <100 μmol/L, or ≥100 μmol/L]) in a 3-arm (1:1:1), double-blind fashion to receive volixibat 20 mg BID, volixibat 80 mg BID, or placebo.
- Supplemental cohort: Up to 30 participants with ICP with screening sBA level from ≥15 to <20 μmol/L will be randomized (based on gestational age at randomization and presence/absence of gestational diabetes) in a 2-arm (1:1), open-label fashion to receive volixibat 20 mg BID or volixibat 80 mg BID.

Study treatment dosing will begin on Day 1 and continue until end of the study treatment period, defined as the day of delivery. Study visits to assess safety, plasma concentrations of volixibat, pharmacodynamics, and efficacy will be conducted weekly from baseline until

delivery. Because alternative measures may be implemented when necessary for public health or other emergency situations, and when sufficient to help ensure the safety of study participants, some visits may be conducted as a home health or remote visit at the discretion of the investigator upon approval by the medical monitor or designee and in accordance with local and regional regulations (see Protocol Appendix 10). Participants will also enter ItchRO assessments into an electronic diary (eDiary) once daily. Use of limited rescue therapies for ICP will be permitted for participants who meet protocol-specified criteria.

A single, formal interim analysis at the end of Part 1 (i.e., after the last study visit for the 60th participant in the primary cohort) that will involve a total of approximately 60 participants (20 per arm) in the primary cohort only will be performed. Data from Part 1 will also be combined with Part 2 and included in the overall safety summaries; however, because this is an operationally seamless adaptive study, efficacy results of participants in Part 1 will not be combined with those from Part 2 for the statistical inference performed in Part 2. Due to the seamless adaptive nature of the study, participants may continue to be accrued after the 60 participants in Part 1 have been enrolled. Under this scenario, participants randomized to placebo and the selected volixibat dose will be used in the analysis in Part 2. Participants randomized to an unselected volixibat dose will not be part of the analysis in Part 2 but will be included within safety summaries.

The interim analysis at the conclusion of Part 1 will be performed by the sponsor and will inform the dose to be explored within Part 2 of the study. Separately, a chartered, independent data monitoring committee (IDMC) will monitor unblinded maternal and fetal safety/plasma drug level data on an ongoing basis during this study (throughout Parts 1 and 2), with the initial IDMC meeting (post-study start) to be held after 12 participants in the primary cohort complete the study; cadence for subsequent meetings will be outlined in the IDMC charter.

Part 2: Confirmation of Safety and Efficacy with Selected Dose

The second stage will constitute a second double-blind, randomized, placebo-controlled phase comparing the selected volixibat dose versus placebo for superior efficacy. For Part 2, it is anticipated that approximately 200 participants (100 per arm) will be randomized.

Study visits to assess safety, pharmacodynamics, efficacy, and plasma concentrations of volixibat in all participants will be conducted weekly from baseline until delivery; sampling for volixibat concentrations may be discontinued in Part 2 if Part 1 results demonstrate consistently negligible or below the limit of quantitation (LOQ) concentrations in maternal and umbilical cord (fetal) plasma. Because alternative measures may be implemented when necessary for public health or other emergency situations and when sufficient to help ensure the safety of study participants, some visits may be conducted as a home health or remote visit at the discretion of the investigator upon approval by the medical monitor or designee and in accordance with local and regional regulations. Participants will also enter ItchRO assessments into an eDiary once daily. Use of limited rescue therapies for ICP will be permitted for participants who meet protocol-specified criteria.

In both Part 1 and Part 2, safety assessments will include AEs, clinical laboratory tests, vital signs, and physical examinations. Additional fetal safety assessments will include serial ultrasounds to assess fetal growth and nonstress tests (NSTs)/cardiotocography (CTG) or

fetal biophysical profile (BPP). Assessment of plasma concentrations of volixibat will be obtained in maternal plasma and umbilical cord plasma (when possible) at single time points, with no formal pharmacokinetic (PK) parameter analysis planned. PD assessments will include bile acids alongside other biomarkers.

The sponsor plans to continue formal safety and efficacy monitoring via a chartered IDMC throughout the study; this will include AE monitoring specifications.

1.3. Sample Size Determination

For the interim analysis at the conclusion of Part 1, 60 participants will have been randomized 1:1:1. Part 1 of the study is designed for estimation without formal statistical testing; hence, there are no formal power considerations. With 20 participants per arm and assuming a (pooled) standard deviation of 25, a 95% confidence interval (CI) for the difference between each volixibat arm and placebo would have a half width of 16.

For the power calculations for Part 2, it is assumed that only a single dose will be selected at the end of Part 1, and a 2-sided alpha 0.05 will be used to test whether there is a difference between the two groups. Thus, for Part 2, a sample size of 100 participants per arm has 80% power to detect an effect size of 10 with standard deviation of 25. Power was calculated using SAS, version 9.4.

Up to 30 additional participants will be enrolled in an open-label supplemental cohort consisting of participants with ICP with screening sBA from ≥15 to <20 μmol/L in a 1:1 manner to the two doses of volixibat.

2. TYPE OF PLANNED ANALYSIS

2.1. Interim Analyses (Part 1)

A single, formal interim analysis will be performed at the end of Part 1 involving a total of approximately 60 participants (20 per arm) in the primary cohort only.

2.2. Supplemental Cohort Analysis

Once the last participant in the supplemental cohort has completed the study, an analysis of the supplemental cohort will be performed. If the supplemental cohort is not fully enrolled by the time of the interim analysis, a separate analysis of the supplemental cohort will be conducted to help inform dose selection for Part 2.

2.3. Final Analysis (Part 2)

A final analysis will be performed after all participants have completed the study. The primary focus of this analysis will be Part 2, but it will include a summary of Part 1, Part 2, and the supplemental cohort. Each of these analyses will be presented separately.

3. GENERAL CONSIDERATIONS FOR DATA ANALYSES

Unless otherwise specified, analyses will be conducted separately on both the primary cohort and the open-label supplemental cohort. Of note, a pooled safety analysis combining the primary and open-label supplemental cohorts may be considered.

All inferential statistical tests will be 2-sided. Summary statistics for continuous variables will include mean, median, standard deviation, standard error of the mean, minimum, maximum, and quartiles. Summary of categorical variables will include sample size, number of occurrences, and percentages (95% CIs will be provided where appropriate).

The study has 3 separate, self-standing cohorts, and each will have its own separate analyses. Primary cohort Parts 1 and 2 will be combined for safety analyses.

Data from Part 1 will also be combined with Part 2 and included in the overall safety summaries; however, since this is an operationally seamless adaptive study, efficacy data from participants in Part 1 will not be combined with those from Part 2 for the statistical inference performed in Part 2. Sampling to assess plasma exposures to volixibat may be discontinued in Part 2 pending results in Part 1.

Analysis results will be presented using descriptive statistics. For categorical variables, the number and percentage of participants in each category will be presented; for continuous variables, the number of participants (n), mean, standard deviation (SD), median, first quartile (Q1), third quartile (Q3), minimum, and maximum will be presented.

By-participant listings will be presented for all participants in the Intent-to-Treat (ITT) Analysis Set and sorted by participants ID number, visit date, and time (if applicable). Data collected on log forms, such as AEs, will be presented in chronological order within each participant. Age, sex at birth, race, and ethnicity will be included in the listings, as space permits.

3.1. Analysis Sets

Analysis sets define the participants to be included in an analysis. Analysis sets and their definitions are provided in this section. The analysis set will be identified and included as a subtitle of each table, figure, and listing.

For each analysis set, the number and percentage of participants eligible for inclusion, as well as the number and percentage of participants who were excluded and the reasons for their exclusion, will be summarized.

A listing of reasons for exclusion from analysis sets will be provided by participants.

The analysis sets are defined in Table 1.

Table 1 Analyses Sets

Analysis Set	Description
Enrolled	All participants who sign the ICF
ITT	All randomized participants
mITT	All randomized participants who receive at least 1 dose of study drug (partial or complete). Participants will be analyzed according to the treatment to which they were randomized.
PP	All participants in the mITT population who do not experience any major protocol violations
Safety	All participants who receive at least 1 dose of study drug. Participants will be classified based upon the treatment they actually receive.

ICF=informed consent form; ITT=intent-to-treat; mITT=modified intent-to-treat; PP=per-protocol.

3.2. Participants Grouping

For analyses based on the ITT and mITT Analysis Sest, participants will be grouped according to the treatment to which they were randomized. For analyses based on the PP and Safety Analysis Sets, participants will be grouped according to the actual treatment received. The actual treatment received will differ from the randomized treatment only when their actual treatment differs from randomized treatment for the entire treatment duration.

3.3. Strata and Covariates

Participants will be randomly assigned to treatment groups via the interactive voice or web response system (IXRS).

- For the primary cohort, a dynamic allocation method, introduced by Pocock and Simon (1975) will be adopted to balance participant assignment between treatment arms based on the following factors: gestational age (<32 weeks or ≥32 weeks) at randomization, presence/absence of gestational diabetes, and highest sBA before randomization (<40 µmol/L, ≥40 µmol/L to <100 µmol/L, or ≥100 µmol/L) after confirmation of study eligibility. Participants with highest pre-randomization sBA <40 µmol/L shall comprise no more than 33% of the primary cohort for Part 1 and Part 2, respectively.
    - For Part 1, in a 1:1:1 ratio via a computer-generated randomization schedule to receive volixibat 20 mg BID, volixibat 80 mg BID, or placebo BID
    - For Part 2, in a 1:1 ratio via a computer-generated randomization schedule to receive the selected volixibat dose (determined at the interim analysis) or placebo BID
- For the supplemental cohort, participants will be randomized with stratification (based on gestational age at randomization and presence/absence of gestational diabetes) after confirmation of study eligibility in a 1:1 ratio via a computer-generated randomization schedule to receive volixibat 20 mg BID or 80 mg BID.

Efficacy endpoints will be evaluated using stratification factors as covariates or stratification variables for analyses, as specified in Section 6. If there is an imbalance in presumed prognostic baseline characteristics between treatment groups, efficacy evaluations may be performed that include these baseline values in efficacy analysis models as covariates; these evaluations will be considered sensitivity analyses.

For efficacy endpoints, the baseline value of the efficacy variable(s) will be included as a covariate in the efficacy analysis model.

3.4. Examination of Participant Subgroups

Subgrouping of participants based on randomization stratification factors will be explored for subgroup analyses. If there is an imbalance between treatment groups in presumed prognostic baseline characteristics that are not stratification factors, subgroupings based on these imbalanced baseline characteristics will also be explored for analysis of the primary and secondary endpoints. The presumed prognostic baseline characteristics include the following:

- Age (≤35 years and >35 years)
- Race/Ethnicity (Hispanic, White [not Hispanic], South Asian [not Hispanic], and Other)
- Baseline ItchRO: (4-7 and >7-10)
- Prior history of ICP (Yes and No)

3.5. Multiple Comparisons

In order to maintain study-wide Type I error control, a hierarchical testing procedure will be used in the comparisons between volixibat and placebo on the primary and secondary efficacy endpoints in the mITT population. The hierarchical order for testing the null hypotheses is as follows:

(1) Mean change in total sBA between baseline and Week 3
(2) Mean change in Adult ItchRO between baseline and Week 3
(3) The proportion of participants experiencing a composite perinatal outcome

The effect of such a procedure is that no confirmatory claims can be based on the endpoint(s) that have a rank lower than that endpoint whose null hypothesis was the first that could not be rejected. Those tests will be considered as exploratory without any formal conclusions drawn about them.

3.6. Missing Data and Outliers

3.6.1. Missing Data

In general, missing data will not be imputed unless methods for handling missing data are specified. Exceptions are presented in this document.

For missing last dosing date of study drug, imputation rules are described in Section 4.2.1. The handling of missing or incomplete dates for prior and concomitant medications is described in Section 4.3, for the primary efficacy endpoint is described in Section 6.1.5, and for AE onset is described in Section 7.1.5.2.

3.6.2. Outliers

Outliers will be identified during the data management and data analysis process. Sensitivity analyses may be conducted to assess the robustness of the results

3.7. Data Handling Conventions and Transformations

In general, age (in years) on the date of first dose study drug will be used for analyses and presentation in listings. If an enrolled participants was not dosed with any study drug, the enrollment date will be used instead of the first dosing date of study drug. For screen failures, the date the informed consent was signed will be used for age calculation. If only the birth year is collected on the CRF, "01 July" will be used for the unknown birth day and month for the purpose of age calculation. If only birth year and month are collected, "01" will be used for the unknown birth day.

Data that are continuous in nature but are less than the lower LOQ or above the upper LOQ will be imputed as follows:

- A value that is 1 unit less than the LOQ will be used to calculate descriptive statistics if the datum is reported in the form of "< x" (where x is considered the LOQ). For example, if the values are reported as <50 and <5.0, values of 49 and 4.9, respectively, will be used to calculate summary statistics. An exception to this rule is any value reported as <1 or <0.1, etc. For values reported as <1 or <0.1, a value of 0.9 or 0.09, respectively, will be used to calculate summary statistics.

- A value that is 1 unit above the LOQ will be used to calculate descriptive statistics if the datum is reported in the form of "> x" (where x is considered the LOQ). Values with decimal points will follow the same logic as above.

- The LOQ will be used to calculate descriptive statistics if the datum is reported in the form of "≤ x" or "≥ x" (where x is considered the LOQ).

If methods based on the assumption that the data are normally distributed are not adequate, analyses may be performed on transformed data or nonparametric analysis methods may be used, as appropriate.

3.8. Analysis Visit Windows

3.8.1. Definition of Study Day

Study day will be calculated from the first dosing date of study drug and derived as follows:

- For postdose study days: Assessment Date – First Dosing Date + 1

- For days prior to the first dose of study drug: Assessment Date – First Dosing Date

Therefore, Study Day 1 is the day of first dose of study drug administration. Please note that this differs by 1 day relative to days specified within the protocol.

3.8.2. Analysis Visit Windows

Participant visits might not occur on protocol-specified days. Therefore, for the purpose of analysis, observations will be assigned to analysis windows. A notable exception to the analysis visit windows below is the end of treatment/date of delivery. This will be summarized as its own visit and that will be the value used for endpoints that look at either the value or change from baseline to delivery.

The analysis windows for ItchRO, sBA, hematology, chemistry, coagulation, weight, vital Signs, and urinalysis are provided in Table 2.

The analysis windows for 5-D Itch, EQ-5D-5L, PROMIS Fatigue and Sleep patient reported outcomes (PROs), PIS-Itch, PGIC-Itch, and Biomarker samples (7α-hydroxy-cholesten-3-one [7αC4], fibroblast growth factor 19 [FGF-19], autotaxin, progesterone sulfate, and glucagon-like peptide 1 [GLP-1]) are provided in Table 3.

The analysis windows for lipid panel and glycated albumin are provided in Table 4.

The analysis windows for vitamin A, D, and E levels and fetal ultrasound are provided in Table 5.

Table 2 Analysis Visit Windows for ItchRO^, sBA, Hematology*, Chemistry*, Coagulation*, Weight, Vital Signs, and Urinalysis

Nominal Visit	Nominal Study Day	Lower Limit	Upper Limit
Baseline	1	(none)	1
Week 1	8	2	11
Week K (K is every week after previous visit)	K*7 + 1	K*7 – 2	K*7 + 4

^ ItchRO is collected daily, but will have weekly averages calculated and results will be grouped according to the analysis windows above.
* Hematology, chemistry, and coagulation are not planned to be collected at the baseline visit. Rather, the screening visit (or a subsequent unscheduled visit) will serve as the baseline.

Table 3 Analysis Visit Windows for 5-D Itch, CSS, EQ-5D-5L, PROMIS Fatigue and Sleep PROs, PIS-Itch, PGIC-Itch, and Biomarker Samples (C4, FGF-19, Autotaxin, Progesterone Sulfate, GLP-1)

Nominal Visit	Nominal Study Day	Lower Limit	Upper Limit
Baseline*	1	(none)	1
Week 3	22	19	25

*PGIC-Itch was not collected at baseline.

Table 4 Analysis Visit Windows for Lipid Panel and Glycated Albumin

Nominal Visit	Nominal Study Day	Lower Limit	Upper Limit
Baseline	1	(none)	1
Week 2	15	2	22
Week K (K is every 2 weeks after previous visit)	K*7 + 1	(K – 1)*7 + 2	(K+1)*7 + 1

Table 5 Analysis Visit Windows for Vitamin A, D, and E Levels and Fetal Ultrasound

Nominal Visit	Nominal Study Day	Lower Limit	Upper Limit
Baseline	1	(none)	1
Week 4	29	2	43
Week K (K is every 4 weeks after previous visit)	K*7 + 1	(K – 2)*7 + 2	(K+2)*7 + 1

3.8.3. Selection of Data in the Event of Multiple Records in an Analysis Visit Window

Depending on the statistical analysis method, single values may be required for each analysis window. For example, change from baseline by visit usually requires a single value, whereas a time-to-event analysis would not require 1 value per analysis window.

If multiple valid, nonmissing, continuous measurements exist in an analysis window, records will be chosen based on the following rules if a single value is needed:

- In general, the baseline value will be the last nonmissing value on or prior to the first dose of study drug, unless specified differently. If multiple measurements occur on the same day, the last nonmissing value prior to the time of first dose of study drug will be considered as the baseline value. If these multiple measurements occur at the same time or the time is not available, the average of these measurements (for continuous data) will be considered the baseline value.

- For postbaseline values:
  - The record closest to the nominal day for that visit will be selected.
  - If there are 2 records that are equidistant from the nominal day, the later record will be selected.
  - If there is more than 1 record on the selected day, the average will be taken, unless otherwise specified.
  - If a participant gives birth prior to Week 3, then the value obtained at the date of delivery will be analyzed according to the next analysis window after the previous visit. For example, if there was a Week 2 sample provided on Day 15 and then the participant delivered on Day 18, the Day 18 record would be mapped to Week 3, even though it is contained within the upper bound of the Week 2 analysis window. This is designed to minimize the amount of missing data for the primary endpoint.

If multiple valid, nonmissing, categorical measurements exist in an analysis window, records will be chosen based on the following rules if a single value is needed:

- For baseline, the last available record on or prior to the date of the first dose of study drug will be selected. If there are multiple records with the same time or no time recorded on the same day, the value with the lowest severity will be selected (e.g., normal will be selected over abnormal for safety electrocardiogram [ECG] findings).
- For postbaseline visits, if there are multiple records with the same time or no time recorded on the same day, the value with the worst severity within the window will be selected (e.g., abnormal will be selected over normal for safety ECG findings).

4. PARTICIPANT DISPOSITION

4.1. Participant Enrollment and Disposition

A summary of participant enrollment will be provided by treatment group for each investigator and overall. The summary will present the number and percentage of participants enrolled. For each column, the denominator for the percentage calculation will be the total number of participants analyzed for that column.

A similar enrollment table will be provided by randomization stratum. The denominator for the percentage of participants in the stratum will be the total number of enrolled participants. If there are discrepancies in the value used for stratification assignment between the IXRS and the clinical database, the value collected in the clinical database will be used for the summary. A listing of participants with discrepancies in the value used for stratification assignment between the IXRS and the clinical database at the time of data finalization will be provided.

The randomization schedule used for the study will be provided as an appendix to the CSR.

A summary of participant disposition will be provided by treatment group and part (Part 1 [double-blind], Part 2 [double-blind], and Open-Label). This summary will present the number of participants screened, the number of participants who met all eligibility criteria but were not randomized with reasons participants not randomized, the number of participants randomized, and the number of participants in each of the categories listed below:

- Safety Analysis Set
- Intent-to-Treat
- Modified Intent-to-Treat
- Per-Protocol Analysis Set
- Continuing study drug
- Completed study drug
- Did not complete study drug with reasons for premature discontinuation of study drug
- Continuing study
- Completed study phase
- Did not complete the study phase with reasons for premature discontinuation of study

For the status of study drug and study completion and reasons for premature discontinuation, the number and percentage of participants in each category will be provided. The denominator for the percentage calculation will be the total number of participants in the Safety Analysis Set corresponding to that column. In addition, a flowchart will be provided to depict the disposition.

The following by-participant listings will be provided by participant identification (ID) number in ascending order to support the above summary tables:

- Reasons for premature study drug or study discontinuation
- Lot number

4.2. Extent of Study Drug Exposure and Adherence

Extent of exposure to study drug will be examined by assessing the total duration of exposure to study drug and the percent of study drug dose received as specified in the protocol.

4.2.1. Duration of Exposure to Study Drug

Total duration of exposure to study drug will be defined as last dosing date minus first dosing date plus 1, regardless of any temporary interruptions in study drug administration except for protocol-specified dose holds and/or dose reductions, and will be expressed in weeks using up to 1 decimal place (eg, 4.5 weeks). If the last study drug dosing date is missing, the latest date among the study drug end date, clinical visit date, laboratory sample collection date, and vital signs assessment date will be used.

The total duration of exposure to study drug will be summarized using descriptive statistics and using the number (ie, cumulative counts) and percentage of participants exposed through the following time periods: 1 day, 3 weeks, 6 weeks, 9 weeks, 12 weeks, 15 weeks, 18 weeks, and >18 weeks. Summaries will be provided by treatment group for the Safety Analysis Set.

No formal statistical testing is planned.

A listing of exposure to the study drug will be provided.

4.2.2. Adherence to Study Drug

The total number of tablets administered will be summarized using descriptive statistics.

The presumed total number of doses tablets administered to a participant will be determined by the data collected within IXRS for drug accountability using the following formula:

$$\text{Total Number of Doses Administered} = \left(\sum \text{No. of Doses Dispensed}\right) - \left(\sum \text{No. of Doses Returned}\right)$$

The level of on-treatment adherence to the study drug regimen will be determined by the total amount of study drug administered relative to the total amount of study drug expected to be administered during a participant's actual on-treatment period based on the study drug regimen.

The level of on-treatment adherence will be expressed as a percentage using the following formula:

$$\text{On-Treatment Adherence (\%)} = \left(\frac{\text{Total Amount of Study Drug Administered}}{\text{Study Drug Expected to be Administered on Treatment}} \right) \text{x } 100$$

Where, the amount of study drug expected to be adminstered on treatment will be derived as 2 x Duration of Exposure to Study Drug, which is defined in Section 4.2.1. Descriptive statistics for the level of on-treatment adherence with the number and percentage of participants belonging to adherence categories (eg, <80%, ≥80 to <90%, ≥90%) will be provided by treatment group for the Safety Analysis Set.

4.3. Prior and Concomitant Medications

Medications collected at screening and during the study will be coded using the current version of the World Health Organization (WHO) Drug dictionary.

4.3.1. Prior Non-Antipruritic Medications

Prior medications are defined as any medications taken prior to the first dose of study drug. Prior antipruritic therapies are described in Section 5.4 and not repeated here.

Prior medications will be summarized by Anatomical Therapeutic Chemical (ATC) drug class Level 2 and preferred name using the number and percentage of participants for each indication. A participant reporting the same medication more than once will be counted only once when calculating the number and percentage of participants who received that medication. The summary will be ordered alphabetically by ATC medical class and then by preferred term in order of descending overall frequency within each ATC medical class. For drugs with the same frequency, sorting will be done alphabetically.

For the purposes of analysis, any medication with a start date prior to the first dosie of study drug will be included in the prior medication summary regardless of when the stop date is. If a partial start date is entered the medication will be considered prior unless the month and year (if day is missing) or year (if day and month are missing) of the start date are after the first dosing date. Medications with a completely missing start date will be included in the prior medication summary, unless otherwise specified.

Summaries will be based on the Safety Analysis Set. No formal statistical testing is planned.

4.3.2. Concomitant Medications

Concomitant medications are defined as medications taken after the first dose of study drug. Use of concomitant medications will be summarized by ATC drug class Level 2 and preferred name using the number and percentage of participants for each indication. A participant reporting the same medication more than once will be counted only once when calculating the number and percentage of participants who received that medication. The summary will be ordered alphabetically by ATC medical class and then by preferred term in descending overall frequency within each ATC medical class. For drugs with the same frequency, sorting will be done alphabetically.

For the purposes of analysis, any medications with a start date prior to or on the first dosing date of study drug and continued to be taken after the first dosing date, or started after the first dosing date but prior to or on the last dosing date of study drug will be considered concomitant medications. Medications started and stopped on the same day as the first dosing date or the last dosing date of study drug will also be considered concomitant. Medications with a stop date prior to the date of first dosing date of study drug or a start date after the last dosing date of study drug will be excluded from the concomitant medication summary. If a partial stop date is entered, any medication with the month and year (if day is missing) or year (if day and month are missing) prior to the date of first study drug administration will be excluded from the concomitant medication summary. If a partial start date is entered, any medication with the month and year (if day is missing) or year (if day and month are missing) after the study drug stop date will be excluded from the concomitant medication summary. Medications with completely missing start and stop dates will be included in the concomitant medication summary, unless otherwise specified. Summaries will be based on the Safety Analysis Set. No formal statistical testing is planned.

All prior and concomitant medications (other than per-protocol study drugs) will be provided in a by-participant listing sorted by treatment group, participant ID and administration date in chronological order.

4.4. Protocol Deviations

Participants who did not meet the eligibility criteria for study entry, but enrolled in the study will be summarized. The summary will present the number and percentage of participants who did not meet at least 1 eligibility criterion and the number of participants who did not meet specific criteria based on the ITT Analysis Set. A by-participant listing will be provided for those participants who did not meet at least 1 eligibility (inclusion or exclusion) criterion. The listing will present the eligibility criterion (or criteria if more than 1 deviation) that participants did not meet and related comments, if collected.

Protocol deviations occurring after participants entered the study are documented during routine monitoring. The number and percentage of participants with important protocol deviations by deviation reason (e.g., nonadherence to study drug, violation of select inclusion/exclusion criteria) will be summarized by indication for the ITT Analysis Set. A by-participant listing will be provided for those participants with important protocol deviations.

5. BASELINE CHARACTERISTICS

5.1. Demographics

Participants demographic variables (i.e., age, sex, race, and ethnicity) will be summarized using descriptive statistics for age, and using number and percentage of participants for sex, race, and ethnicity. The summary of demographic data will be provided for the ITT Analysis Set.

A by-participant demographic listing, including the informed consent date, will be provided by participant ID number in ascending order.

5.2. Other Baseline Characteristics

Other baseline characteristics include height (cm), weight (kg), baseline PRO assessments (including 5-D Itch, CSS, EQ-5D-5L, PROMIS Fatigue & Sleep PROs, and PIS-Itch) and baseline levels of biochemical markers of cholestatis and liver disease (including total sBA, ALP, AST, ALT, bilirubin [total and direct], FGF-19, autotaxin, 7αC4, progesterone sulfate, and GLP-1).

Additional disease-specific variables are discussed in Section 5.4. These baseline characteristics will be summarized using descriptive statistics for continuous variables and using number and percentage of participants for categorical variables. The summary of baseline characteristics will be provided for the ITT Analysis Set. No formal statistical testing is planned.

A by-participant listing of other baseline characteristics will be provided by participant ID number in ascending order.

5.3. Medical History

Medical history will be collected at screening for disease-specific and general conditions (i.e., conditions not specific to the disease being studied) and coded using Medical Dictionary for Regulatory Activities (MedDRA) version 23.1.

A summary of disease-specific medical history will be provided for the ITT Analysis Set.

Time since ICP diagnosis (weeks) will be calculated by (date of informed consent of study drug – date of ICP diagnosis)/7. Time since ICP diagnosis will be summarized using summary statistics for a continuous variable. The other variables will be presented using summary statistics for a categorical variable. No formal statistical testing is planned.

Menstrual history (e.g., last menstrual period [LMP] and date of positive human chorionic gonadotropin test), and obstetric history (e.g., previous pregnancies including terminations of pregnancy and/or spontaneous abortions, genetic screenings, aneuploidy screening, teratogen exposures since LMP/pregnancy, infections [i.e., Covid-19, Zika, HIV, or hepatitis], travel

outside the country, history of IVF or other fertility treatments, history of cerclage placement) will be obtained.

General medical history data will not be summarized, but will be listed.

A by-participant listing of disease-specific and general medical history will be provided by participant ID number in ascending order.

In deriving the time since ICP diagnosis, all partial dates of diagnosis will be identified, and the partial dates will be imputed as follows:

- If day and month are missing but year is available, then the imputed day and month will be 01 Jan.
- If day is missing but the month and year are available, then the imputed day will be the first day of the month.
- Partial date will not be imputed if the year is missing.

5.4. Prior Antipruritic or Prior ICP Therapy

The number of prior antipruritic and prior ICP therapies will be coded using the international reference for medicinal product information (WHODrug) 2020 March version or later and summarized by ATC drug class Level 2 and preferred name using the number and percentage of participants for each indication using descriptive statistics based on the ITT Analysis Set.

6. EFFICACY ANALYSES

6.1. Primary Efficacy Endpoint

6.1.1. Definition of the Primary Efficacy Endpoint

The primary efficacy endpoint analysis is defined as the mean change in total sBA concentration from baseline to Week 3.

6.1.2. Estimand and Analysis of the Treatment Effect

The estimand for the treatment effect will be the mean difference between volixibat and placebo for the change from baseline in sBA at Week 3.

The primary analysis will be conducted in the primary cohort using the modified ITT (mITT) population. A restricted maximum likelihood (REML)–based mixed-effects model for repeated measures (MMRM) will be used including weekly postbaseline change from based line in total sBA during the core study period (double-blind treatment period) through Week 3, as the dependent variable and the fixed, continuous effects of baseline sBA and gestational age, fixed categorical effects of presence/absence of gestational diabetes, treatment group, visit, and treatment group by-visit interaction as covariates and participant as a random effect.

The unstructured variance/covariance matrix will be used to model the variances and covariances for the time points included in the model. The unstructured variance/covariance does not impose any restrictions on the pattern of the matrix elements. The use of an endpoint 3 weeks after baseline should result in minimal amounts of missing data. Every attempt (e.g., relaxing the convergence criteria, increasing the iteration limit, choosing reasonable starting values for the estimates) will be made to ensure convergence using the unstructured modeling of within-participant correlations. However, if the numerical algorithm for estimation of the mixed model fails to converge, the following variance/covariance matrix structures will be used in the following order: 1) heterogeneous Toeplitz, 2) heterogeneous compound symmetry, and 3) heterogeneous autoregressive of order 1. The first (co)variance structure that does not have a convergence problem will be the one used for the analysis. The Kenward-Roger approximation (1997) will be used to estimate denominator degrees of freedom.

The primary efficacy analysis will compare volixibat and placebo using the contrast (difference in least squares [LS] means) between treatment groups at Week 3. Significance tests will be based on LS means using a 2-sided significance level (2-sided 95% CIs).

The null hypothesis for the primary efficacy endpoint of the equality of volixibat and placebo is:

H_0: mean change in total sBA baseline and Week 3 in the two treatment groups are equal

The null hypothesis of equal treatment effect will be rejected if the statistical analysis results in a 2-sided p-value for treatment at Week 3 is ≤0.05. For each cohort, LS means will be calculated for each treatment group for each postbaseline visit in the model. The difference between volixibat and placebo change from baseline in sBA will be estimated, with the corresponding 2-sided 95% CI constructed for each visit through Week 3. The change from baseline LS means with standard error, 95% CI for the LS means, p-value for testing if the LS mean is zero, LS mean difference between treatment groups (volixibat minus placebo) with standard error, 95% CI for the LS mean difference, and p-value for testing if the treatment LS means are equal will be presented.

The study will be claimed successful if the hypothesis of no treatment effect on the primary efficacy endpoint over the primary cohort in the mITT population from Part 2 is rejected at the 0.05 (2-sided) significance level.

Summary statistics on total sBA by treatment group and visit will also be presented. In addition, a figure with the LS mean ± standard error of change from baseline in total sBA will be presented by treatment group and visit.

In addition, a responder analysis (based on change in total sBA) may be considered.

Separately, a similar analysis on the open-label supplemental cohort will be performed. The analysis will be identical, except that baseline sBA will not be a covariate used within the model.

6.1.3. Intercurrent Events and Handling Strategies

The following table listed outs some potential intercurrent events that may impact the estimation of the primary estimand of the study. For each type intercurrent event, we proposed handling strategies:

Events	Potential Consequences	Handling Strategy
Early discontinuation and/or missing because of: 1. Lack of efficacy; physician decision; or required prohibited medication; 2.Adverse events or death; 3. Noncompliance to study drug; 4. Other reasons for treatment discontinuations	• Confounded treatment effect	Hypothetical Policy • Sensitivity Analysis: – MAR – MNAR

Concomitant Medications	• More difficult to pinpoint
	  ○ Administered as needed for both active treatment and placebo
	  ○ May lead to various impacts on participants' treatment/safety/efficacy
		Randomization/Treatment Policy

Other Intrinsic/extrinsic factors • Stratification factors • Age • Sex • Race/Ethnicity	• Differentiated treatment effects	Principal Stratum Policy • Stratified analyses • Subgroup analyses • Covariate adjusted analyses

6.1.4. Rationale for Primary Analysis Method

The MMRM method has been demonstrated extensively as an appropriate choice for the primary analysis in longitudinal confirmatory clinical trials with continuous endpoints (Mallinckrodt et al. 2008). This analysis method, which is from a broader class of direct-likelihood analyses methods, makes use of fully and partially observed data sequences from individual participants by estimating the covariance between data from different time points (Molenberghs and Kenward 2007). Further, it is often useful to implement MMRM using an unstructured approach to modeling both the treatment-by-time means and the (co)variances, leading to what is essentially a multivariate normal model wherein treatment group means at the primary time point are adjusted to reflect both the actual observed data and the projected outcomes from the participants with missing data (Cnaan et al. 1997; Molenberghs et al. 2004; Molenberghs and Kenward 2007).

Despite careful planning and study conduct, the occurrence of missing data cannot be completely eliminated. As a direct likelihood method, the MMRM method is a preferred approach for handling missing data in such designs. MMRM is a full multivariate model in nature that avoids potential bias as a predetermined model and operates in a more general missing at random (MAR) framework (Mallinckrodt et al. 2001). Data are considered MAR if, conditional upon the independent variables in the analytic model, the missingness depends on the observed outcomes of the variable being analyzed but does not depend on the unobserved outcomes of the variable being analyzed. This assumption implies that the behavior of the post dropout observations can be predicted from the observed variables, and therefore that treatment effect can be estimated without bias using the observed data (European Medicines Agency 2010). For studies of missing data in a controlled clinical trial setting, MAR is usually considered as a plausible underlying missing mechanism (Molenberghs and Kenward 2007; Siddiqui et al. 2009; Mallinckrodt et al. 2008, Mallinckrodt et al. 2013). The assumption of MAR is often reasonable because, particularly in longitudinal studies wherein the evolution of treatment effects is assessed by design over time, the observed data and the models used to analyze them can explain much of the missingness (Little and Rubin 1987; Verbeke and Molenberghs 2000). This point may be especially relevant in well-controlled studies, in which extensive efforts are made to observe

all outcomes and factors that influence them while participants are following protocol-defined procedures. Thus, longitudinal clinical trials by their very design aim to reduce the amount of missing not at random (MNAR) data (missingness explained by unobserved responses), thereby increasing the plausibility of MAR (Mallinckrodt et al. 2008).

6.1.5. Sensitivity Analysis of the Primary Efficacy Endpoint

Sensitivity and supportive analysis will be performed on the primary efficacy endpoint to quantify the possible impact of missing data, intercurrent events and to demonstrate the robustness of the conclusions. See below for details about missing data handling, placebo-based imputation, and a tipping-point analysis.

Furthermore, the primary analysis will be repeated using a hypothetical (de jure) estimand where data collected after rescue medication is initiated will be omitted. This can be viewed as addressing a per protocol estimand, or more simply the applicable estimand if all participants completed the first 3 weeks of follow-up without any rescue or without any missing data.

6.1.6. Missing Data

Although the assumption of MAR, as used for the primary efficacy analysis method, is often reasonable in clinical studies, the possibility of MNAR data cannot be ruled out. Therefore, examining the robustness of the primary analysis results under MNAR is desired. Both a MNAR- and MAR-based analyses will be the basis upon which sensitivity of missing data is assessed.

Sensitivity analyses to deal with missing data and intercurrent events will be based on multiple imputation methods where missing values are imputed individually under different assumptions. Multiple imputation is a simulation-based approach where missing values are replaced using an appropriate stochastic model given the observed data and covariates, creating multiple completed data sets. These completed datasets are then analyzed using standard analysis methods (i.e., MMRM), and the different parameter estimates across the datasets are then combined to produce a unique point estimate and standard error taking into account the uncertainty of the imputation process. The following 2 sensitivity analysis models, one based on the standard MAR approach and the other based on the MNAR approach, will be used to examine robustness of the primary analysis results under the impact of missing data and intercurrent events.

6.1.7. Complete Multiple Imputation Differentiating Interrcurrent Events

For participants with complete data up to a particular visit, a distribution will be assumed for the assessments includes the outcome at that visit based on predefined missing categories and types of intercurrent events. The missing data then will be imputed based on specified algorithm which draws samples based on the prespecified distribution and algorithm. This process will be repeated a given number of times, resulting in the same number of complete analysis data sets. The MMRM analyses then will be performed separately for each of the completed analysis data sets, and the results will be combined into one multiple imputation inference (Little and Yau 1996; Schafer 1997). Example SAS code is provided in Appendix 3.

Imputation distribution:

The imputation distribution for the missing change from baseline in sBA at visit *t* will be assumed to follow a normal distribution. All randomized participants will be classified in one of the three categories based on the following rules: non-missing category, missing category 1, or missing category 2,

Non-missing category: Participants who have completed sBA assessment at baseline and Week 3 and have the change from baseline Week 3.

Missing category 1: Participants who discontinued treatment because of adverse events, lack of efficacy, noncompliance with study drug, death, physician decision, or required prohibited medication. These participants have missing change from baseline in sBA at visits including Week 3.

Missing category 2: Participants who discontinued treatment because of all other reasons other than the ones listed in category 1 and have missing the change from baseline in sBA at visits including Week 3, or participants who have completed 3 weeks treatment duration but are missing the change from baseline in sBA at visits including Week 3.

Imputation algorithm:

We use the following algorithm that relates the mean of the missing change from baseline in sBA at visit *t* to the missing categories defined above. The algorithm will be implemented independently within each treatment group:

- **Missing category 1:** randomly draw a sample from the normal distribution $N(\mu_{25}, \sigma^2)$, where μ_{25} is the 25th percentile of the non-missing change from baseline in sBA at visit *t* and σ^2 is the sample variance estimated using the non-missing change from baseline at visit *t*.

- **Missing category 2:** randomly draw a sample from the normal distribution $N(\mu, \sigma^2)$, where μ is the mean of the non-missing change from baseline in sBA at visit *t* and σ^2 is the sample variance estimated using the non-missing change from baseline at visit *t*.

Analysis model:
The complete MI method is described below:

- Impute missing values using the normal distribution specified in the above algorithm to impute values for all participants visit *t*. After imputation, all participants in the defined analysis population(s) will have non-missing change from baseline in sBA at visit *t*.

- Repeat the imputation for visits *t+1*, …, Week 3, until measurements at visits from visit *t* to Week 3 are imputed to form a complete dataset.

- Repeat the process K (K=5) times, using the procedure described above to form K imputed complete datasets.
- Fit the MMRM model including terms for treatment, and all other terms as in the primary analysis model to each imputed dataset, to estimate the change from baseline during at Week 3 and its variance.
- Combine the results from the K imputed datasets using the SAS procedure MIANALYZE to derive the MI estimator.

We fit the analysis model (MMRM model specified before) to the *kth* imputed dataset, denoting the corresponding estimate of the variance V_k. The MI estimator, $\tilde{\theta}_{MI}$, is the average of the K individual estimators.

The estimated variance of $\tilde{\theta}_{MI}$ is a combination of the between- and within-imputation variability as follows: $V_{MI} = W + \left(1 + \frac{1}{K}\right)$, where $W = \frac{1}{K}\sum_{k=1}^{K} V_k$ is the within-imputation variability and is $B = \frac{1}{K-1}\sum_{k=1}^{K}(\tilde{\theta}_k - \tilde{\theta}_{MI})^2$ is the between-imputation variance. The statistic $T = \frac{\tilde{\theta}_{MI}-\theta}{\sqrt{V_{MI}}}$ has an approximate t_V distribution

6.1.8. Tipping Point Analysis

An additional sensitivity analysis using the tipping-point approach will be conducted to assess the robustness of the primary analysis approach. If it is plausible that, for the active treatment group, the distribution of missing primary endpoint responses has a smaller expected reduction than that of the corresponding distribution of the observed primary endpoint responses, the conclusion under the MAR assumption should be examined. It is desired to impose a fixed and definite set of quantities to encapsulate the change in efficacy associated with withdrawal (missing) for the active treatment group, and independently for the placebo group. The tipping point multiple imputation analysis as described by Ratitch et al. (2013) will be applied. Tipping point analysis is a means of exploring the influence of missingness on the overall conclusion from statistical inference by positing a wide spectrum of specifications regarding the missingness mechanism (from less pessimistic to more pessimistic). The analysis finds a (tipping) point in this spectrum of specifications, at which conclusions change from being favorable to the experimental treatment to being neutral. The analysis will be based on the following procedure:

1. The missing data are filled in m (m = 10) times to generate m complete data sets assuming that we do not differentiate missing data category (similar to assuming the algorithm for missing category 2).
2. The m complete data sets are analyzed by using standard procedures.
3. The results from the m complete data sets are combined for the inference.

4. Repeat the steps #1 to generate multiple imputed data sets, with a specified shift parameter that adjust the imputed values for observations in the relevant treatment group independently.
5. Repeat the step 2 for the imputed data sets with shift parameter applied.
6. Repeat the step 3 to obtain the p-value to see if the p-value is still ≤0.05.
7. Repeat the steps 4-6 with more stringent shift parameter applied until the p-value >0.05.

After such a tipping point is determined, clinical judgment can be applied as to the plausibility of the specifications underlying this tipping point. The tipping point can be identified while the result is no longer statistically significant. This imputation analysis uses a specified sequence of shift parameters, which adjust the imputed values for observations in the active treatment group and independently for the placebo group. A sample of SAS code for sensitivity analysis using the tipping-point method is provided in Appendix 3.

6.2. Secondary Efficacy Endpoints

The secondary efficacy endpoints (see Section 1.1.2) are designed to evaluate the efficacy of volixibat versus placebo for the treatment of pruritus in participants with ICP as well as composite perinatal outcomes.

1. The secondary efficacy endpoint for pruritus is defined as the mean change from baseline to Week 3 in the weekly average worst daily itch score as measured by the Adult ItchRO. The worst daily itch score is averaged over the 7 days preceding baseline and over the 7 days preceding each week through Week 3. If more than 3 days of assessments are missing from a given week, then that weekly average will be set to missing. The analysis of the Adult ItchRO will use an MMRM with a comparable model to that for the primary endpoint.
2. The secondary efficacy endpoint for composite perinatal outcomes is defined as the proportion of participants that experience any of the 4 outcomes specified in Section 1.1.2. For this responder-type endpoint, the number and proportion of participants that are considered a "responder" will be summarized by treatment group at Week 3 and analyzed using a stratified Cochran-Mantel-Haenszel (CMH) test to calculate the p-value for the difference between treatment groups. The p-value from the CMH test will be the basis of the hypothesis testing.

   Additionally, the baseline stratum weighted difference in the response rate ($P_1 - P_2$) and its 95% CI will be calculated based on stratum-adjusted Mantel-Haenszel (MH) proportions as described as follows (Koch 1989), where stratification factors include gestational age, baseline ItchRO score, and gestational diabetes:

$$P_1 - P_2 \pm Z_{(1-\alpha/2)} * SE(P_1 - P_2)$$

   where

- $(P_1 - P_2) = \frac{\sum w_h d_h}{\sum w_h}$ is the stratum-adjusted MH proportion difference, where $d_h = p_{1h} - p_{2h}$ is the difference in the response rate between treatment groups 1 and 2 in stratum h (h=1 to 12).

- $w_h = \frac{n_{1h} n_{2h}}{n_{1h}+n_{2h}}$ is the weight based on the harmonic mean of the sample size per treatment group for each stratum where n_{1h} and n_{2h} are the sample sizes of the treatment groups 1 and 2 in stratum h.

- $SE(P_1 - P_2) = \sqrt{\frac{\sum w_h^2 \left[\frac{p^*_{1h}(1-p^*_{1h})}{n_{1h}-1} + \frac{p^*_{2h}(1-p^*_{2h})}{n_{2h}-1}\right]}{(\sum w_h)^2}}$, where $p^*_{1h} = \frac{m_{1h}+0.5}{n_{1h}+1}$ and $p^*_{2h} = \frac{m_{2h}+0.5}{n_{2h}+1}$. m_{1h} and m_{2h} are the number of participants who are "responders" in treatment groups 1 and 2 in stratum h.

- $\alpha = 0.05$

- $Z_{(1-\alpha/2)} = Z_{0.975} = 1.95996$ is the 97.5th percentile of the normal distribution

  Note that if the computed lower confidence bound is less than -1, the lower bound is defined as -1. Similarly, if the computed upper confidence bound is greater than 1, the upper bound is defined as 1.Furthermore, as supportive analyses, a Logistic Regression (LR) estimator approach (Ge et al. 2011), as well as a logistic regression model with baseline sBA, gestational age, and presence/absence of gestational diabetes as covariates will be run. The model assumptions of the logistic regression model are not considered a concern because a randomized study has comparable distributions of the treatment groups for covariates and factors for stratification, although they may impact the interpretation of estimated parameters for the covariates. Most importantly, a logistic regression model addresses estimands that are conditional on the covariates in the model, with these estimands sometimes being called participant-specific, and express the extent of efficacy of a treatment that a participant might expect on the basis of her covariates, although they tend to be further from the null for an effective treatment than population average estimand. A randomization-based version of logistic regression will also be considered. It is robust to model assumptions and it has a population average estimand that is the same as the unadjusted odds ratio for comparing two treatments (Zink and Koch 2012).

6.3. Exploratory Efficacy Endpoints

The exploratory efficacy endpoints are defined in Section 1.1.3.

Continuous endpoints will be summarized using descriptive statistics. A p-value will be computed for between treatment comparisons using a van Elteren (vE) test (van Elteren 1960). The Wilcoxon–Mann–Whitney (WMW) test (Mann and Whitney 1947; Wilcoxon 1945) is widely used nonparametric approach to compare two treatments. In the

presence of stratum factors such as gestational age or ItchRO(Obs) Score, the van Elteren test, a stratified WMW test, can be used to adjust for the stratification factors. It was shown by Qu et al. (2008) that vE test preserves the type I error rate regardless of the existence of the stratum effects, and is better than the WMW test when the stratum effects are large.

A non-strata adjusted 95% CI for the difference in means between treatment groups will also be presented. A test of equality of variances will be performed. If the null hypothesis that the variances are equal is rejected, then the 95% CI will be constructed using a Satterthwaite approximation. Otherwise, the variances will be pooled.

For categorical endpoints, the number and percent for each category will be displayed. For responder-type analyses, within a treatment group, an approximate 95% CI will be presented using a stratified Wilson confidence interval (Yan and Su 2010). A 95% CI comparing two treatment groups based on stratum-adjusted MH proportions as described in Section 6.2 will also be generated. A stratified CMH test will be used to calculate the p-value for the difference between treatment groups. The method is applicable to the following endpoints:

- Proportion of participants with an early term delivery (37–38 weeks 6 days, inclusive)
- Proportion of participants with a full-term delivery (≥39 weeks)
- Proportion of participants experiencing a perinatal death
- Proportion of participants with a spontaneous preterm delivery
- Proportion of participants with an iatrogenic preterm delivery attributable to ICP
- Proportion of neonates requiring neonatal intensive care unit (NICU) admission for ≥12 hours between birth and hospital discharge
- Proportion of neonates with meconium-stained amniotic fluid at delivery
- Proportion of participants requiring rescue medication for ICP
- Proportion of neonates with Apgar score <7 at 5 minutes of life
- Proportion of neonates with umbilical cord pH <7.0 at birth

"Responder" analyses for composite perinatal outcomes will be performed using baseline sBA and ItchRO and the change at Week 3 values in sBA and ItchRO as covariates and, separately, using the Week 3 values themselves (instead of the change at Week 3).

The change in total sBA from Week 3 (primary endpoint) to delivery will be considered, and the difference between treatment groups will be tested using a randomization-based version of rank analysis of covariance (ANCOVA) (Stokes et al. 2012). In a similar manner, the change from baseline in total sBA concentration to the time of delivery will also be explored. An additional analysis will be performed that examines the change from baseline to delivery but that averages across the last two time points on or prior to delivery. Similar analyses will be performed for Adult ItchRO. Associations between the changes from baseline to Week 3 in total sBA and Adult ItchRO will be assessed using Pearson's and Spearman's correlation coefficient. Additionally, a scatterplot of two variables will also be produced.

The following endpoints will be summarized by treatment arm and the summary statistics by treatment arm and the treatment differences will be tabulated:

- Mean gestational age at delivery
- Mean birth weight percentile
- Mean placental weight percentile
- Means for umbilical cord blood sBA, total cholesterol, LDL cholesterol, and glucose
- Mean maternal estimated blood loss at delivery
- Mean time from randomization to delivery

Further exploratory efficacy endpoints include continuous and binary endpoints on additional perinatal outcomes. The analysis of continuous endpoints will follow the same approach as other exploratory efficacy endpoints. Binary outcomes will be analyzed in the same manner as the secondary efficacy endpoint for composite perinatal outcomes.

The secondary endpoint of composite perinatal outcomes will also be explored using an ordinal scale, with ranks assigned according to the severity of the outcomes specified within Section 3.2.

Table 6 Perinatal Outcome Severity Scoring

Perinatal Outcome	Score
Perinatal death	4
Spontaneous preterm delivery	3
Iatrogenic preterm delivery attribute to ICP or ICP-related complications	2
Neonatal unit admission for ≥12 hours from birth until hospital discharge for any of the reasons listed in Section 1.1.3	1
No perinatal outcome	0

The score is assigned according to the most severe outcome following the order shown above. If a participant experiences perinatal death, spontaneous preterm delivery, or iatrogenic preterm delivery attributed to ICP or ICP-related complications AND neonatal unit admission for ≥12 hours from birth until hospital discharge for any of the reasons listed in Section 3.2, the score will be adjusted by 0.5. For example, if a participant has spontaneous preterm delivery followed by a neonatal unit admission, the score would be 3+0.5=3.5, where the 3 comes from the spontaneous preterm delivery and the 0.5 comes from the fact that neonatal unit admission occurred. This will be analyzed using a stratified log rank test that provides more weight to the more severe outcomes and will employ a randomization-based covariance adjustment (Zink and Koch 2012).

Change in biomarkers of bile acid synthesis, inflammation, and pruritogens as well as the proportion of participants with sBA <40 µmol/L at the end of the treatment period will also be explored. The change in biomarkers will follow the same approach as other exploratory efficacy endpoints described above. The proportion of participants with sBA <40 µmol/L at the end of the treatment period will be analyzed in the same manner as the secondary efficacy endpoint for composite perinatal outcomes.

7. SAFETY ANALYSES

Safety summaries will be separated by the portion of the study in which they occurred. There are two parts of the double-blind part of the study and the open-label supplemental cohort.

7.1. Adverse Events and Deaths

Adverse events are captured both for the participant (mother) and their child. As a result, summaries and listings will be presented separately.

7.1.1. Adverse Event Dictionary

Clinical and laboratory AEs will be coded using the MedDRA version 23.1. System organ class (SOC), high-level group term (HLGT), high-level term (HLT), preferred term (PT), and lower-level term (LLT) will be provided in the AE dataset.

7.1.2. Adverse Event Severity

Adverse events are graded by the investigator as Grade 1, 2, 3, 4, or 5 according to Common Terminology Criteria for Adverse Events (CTCAE) Version 5.0. The severity grade of events for which the investigator did not record severity will be categorized as "missing" for tabular summaries and data listings. The missing category will be listed last in summary presentation.

7.1.3. Relationship of Adverse Events to Study Drug

Related AEs are those for which the investigator selected "Related" on the AE CRF to the question of "Relationship to Study Drug(s)." Relatedness will always default to the investigator's choice, not that of the medical monitor. Events for which the investigator did not record relationship to study drug will be considered related to study drug for summary purposes. However, by-participant data listings will show the relationship as missing.

7.1.4. Serious Adverse Events

Serious adverse events (SAEs) will be identified and captured as SAEs if the AEs met the definitions of SAEs that were specified in the study protocol. SAEs captured and stored in the clinical database will be reconciled with the safety database before data finalization.

7.1.5. Treatment-Emergent Adverse Events

7.1.5.1. Definition of Treatment-Emergent Adverse Events

Maternal treatment-emergent adverse events (TEAEs) are defined as AEs that start or deteriorate on or after the first dose of study medication and no later than 7 days following the last dose of study medication. For any participants (i.e., mothers) who die during the study and the date of death is between the date of the first dose of study medication and the date of study discontinuation (as entered by the site), inclusive, all AEs (including those

resulting in death) that occur during the study will be considered as TEAEs irrespective of the last dose and will be included in the TEAE summaries. All summaries of AEs will be based on TEAEs unless specified otherwise.

For newborns, all reported AEs will be considered TEAEs for the purpose of summaries.

7.1.5.2. Incomplete Dates

If the onset date of the AE is incomplete and the AE stop date is not prior to the first dose of study drug, then the month and year (or year alone if month is not recorded) of onset determine whether an AE is treatment emergent. The event is considered treatment emergent if both of the following 2 criteria are met:

- The AE onset is the same as or after the month and year (or year) of the first dosing date of study drug, and
- The AE onset date is the same as or before the month and year (or year) of the date corresponding to 7 days after the date of the last dose of study drug

An AE with completely missing onset and stop dates, or with the onset date missing and a stop date later than the first dosing date of study drug, will be considered to be treatment emergent. In addition, an AE with the onset date missing and incomplete stop date with the same or later month and year (or year alone if month is not recorded) as the first dosing date of study drug will be considered treatment emergent.

7.1.6. Summaries of Adverse Events and Deaths

Treatment-emergent AEs will be summarized based on the Safety Analysis Set.

7.1.6.1. Summaries of AE Incidence

The number and percentage of participants who experienced AEs described below will be provided and summarized by SOC and PT unless otherwise indicated:

- TEAEs
- TEAEs of Grade 3 or higher (by SOC, PT and maximum severity)
- Treatment-related TEAEs (TRAEs)
- TRAEs of Grade 3 or higher (by SOC, PT and maximum severity)
- Serious TEAEs
- Serious TRAEs
- TEAEs leading to discontinuation of study drug

- TEAEs leading to death
- AESIs
- ECIs

A brief, high-level summary of AEs described above will be provided by the number and percentage of participants who experienced the above AEs.

Multiple events will be counted only once per participant in each summary. Adverse events will be summarized and listed first in alphabetic order of SOC and then by PT in descending order of total frequency within each SOC. For summaries by severity, the most severe severity will be used for those AEs that occurred more than once in a given participant during the study.

In addition to the above summary tables, TEAEs, TEAEs of Grade 3 or higher, TRAEs, TRAEs of Grade 3 or higher, serious TEAEs, TEAEs leading to discontinuation of study drug, AESIs, and ECIs will be summarized by PT only in descending order of total frequency within indication.

Furthermore, data listings will be provided for the following:

- All AEs, indicating whether the event is treatment emergent
- All AEs of Grade 3 or higher
- TRAEs
- TRAEs of Grade 3 or higher
- AESIs
- ECIs
- SAEs
- AEs leading to death (ie, outcome of death)
- AEs leading to discontinuation of study drug

If there are no participants who meet the criteria for any of these listings, the listing will be generated and will contain a statement, "No <*criteria*> were reported." For example, if there are no AEs leading to death on study, the statement would be "No AEs leading to death were reported."

7.1.7. Additional Analysis of Adverse Events

Furthermore, exposure adjusted incidence rates (EAIR) will be provided for each AE reported. For EAIR analyses, the following convention will be used to calculate the denominator (i.e., population-level time at-risk) for these analyses. The population-level time at-risk includes only the time each participant is at risk. The start date of at-risk time is the date of first dose of study drug, and the last date of at-risk time is the start date of the occurrence of the AE or is censored at completion/early termination, whichever is earlier.

While no formal hypothesis testing or margin selection is planned, the risk difference and corresponding exact 95% CI (using the method of Chan and Zhang [1999]) will be presented to compare treatment groups with regards to AESIs and ECIs.

The TEAE by PT summary will also be repeated by the subgroups defined in Section 3.4.

Incidence/prevalence plots of common adverse events may be explored.

7.2. Laboratory Evaluations

Laboratory data collected during the study will be analyzed and summarized using both quantitative and qualitative methods. Summaries of laboratory data will be provided for the Safety Analysis Set and will include data collected up to the last dose of study drug plus 7 days for participants who have permanently discontinued study drug, or all available data at the time of the database snapshot for participants who were ongoing at the time of an interim analysis. The analysis will be based on values reported in conventional units. When values are below the LOQ, they will be listed as such, and the closest imputed value will be used for the purpose of calculating summary statistics as specified in Section 3.7. Hemolyzed test results will not be included in the analysis, but they will be listed in by-participant laboratory listings.

A by-participant listing for laboratory test results will be provided by participant ID number and time point in chronological order for hematology, serum chemistry, and urinalysis separately.

No formal statistical testing is planned.

7.2.1. Summaries of Numeric Laboratory Results

Descriptive statistics will be provided by indication for each laboratory test specified in the study protocol as follows:

- Baseline values
- Values at each postbaseline time point
- Change and percentage change from baseline at each postbaseline time point

A baseline laboratory value will be defined as the last measurement obtained on or prior to the date/time of first dose of study drug. Change from baseline to a postbaseline visit will be

defined as the visit value minus the baseline value. The mean, median, Q1, Q3, minimum, and maximum values will be displayed to the reported number of digits; SD values will be displayed to the reported number of digits plus 1.

7.2.2. Shifts Relative to the Baseline Value

For all laboratory tests, shift tables will be presented by showing changes in results from baseline value (low, normal, and high) to the worst postbaseline result (low, normal, high, or low and high). A participant would have a worst postbaseline result of normal if all postbaseline results are within the normal reference range. The value of "low and high" would be used if the participant has postbaseline values that are below the normal reference range at one visit and above the normal reference range at another visit.

7.3. Body Weight, Height, and Vital Signs

Descriptive statistics will be provided indication for body weight, height, body mass index (BMI), and vital signs as follows:

- Baseline value
- Values at each postbaseline time point
- Postbaseline minimum value
- Change from baseline at each postbaseline time point

A baseline value will be defined as the last available value collected on or prior to the date/time of first dose of study drug. Change from baseline to a postbaseline visit will be defined as the postbaseline value minus the baseline value. Body weight and vital signs measured at unscheduled visits will be included for the baseline value selection.

In the case of multiple values in an analysis window, data will be selected for analysis as described in Section 3.8.3. No formal statistical testing is planned.

A by-participant listing of vital signs will be provided by participant ID number and time point in chronological order. Body weight, height, and BMI will be included in the vital signs listing, if space permits. If not, they will be provided separately.

7.4. Electrocardiogram Results

ECG data will not be presented in the CSR since ECGs will not be assessed in this study other than as part of the screening process for potential new participants.

7.5. Other Safety Measures

7.5.1. Health Utilization Assessments

Healthcare resource utilization variables include the number and duration of medical care encounters, including surgeries and other selected procedures (inpatient and outpatient), the dates and duration of hospitalization (total days or length of stay, including duration by wards [e.g., intensive care unit]), the number and type of diagnostic and therapeutic tests and procedures, and outpatient medical encounters and treatments (including physician or emergency department visits, tests and procedures, and medications). Descriptive statistics including number of observations, mean, 95% CI on the mean, median, minimum, and maximum will be presented by visit for continuous variables. For categorical variables, the number and proportion of participants will be presented.

7.5.2. Newborn Safety Assessments

The number and percentage of participants who experienced AEs described below will be provided and summarized by SOC and PT unless otherwise indicated:

- TEAEs
- TEAEs of Grade 3 or higher (by SOC, PT and maximum severity)
- Serious TEAEs

Furthermore, all AEs reported within newborns will be listed separately.

7.6. Changes From Protocol-Specified Safety Analyses

There are no deviations from the protocol-specified safety analyses.

8. PATIENT-REPORTED OUTCOMES

To evaluate QoL in participants with ICP-associated pruritus treated with volixibat versus placebo on the basis of the following endpoints:

- Change in 5-D Itch scale with 1-week recall from baseline to Week 3 and delivery
- Change in Clinician Scratch Scale (CSS) from baseline to Week 3 and delivery
- Change in EQ-5D-5L from baseline to Week 3 and delivery
- Change in PROMIS fatigue score from baseline to Week 3 and delivery
- Change in PROMIS sleep score from baseline to Week 3 and delivery
- Change in Patient Impression of Severity of Itch (PIS-Itch) from baseline to Week 3 and delivery
- PGIC-Itch at Week 3 and delivery

Because the PROs are administered only at baseline, Week 3, and end of treatment, the change from baseline will be compared between the two groups with use of rank ANCOVA with adjustment for baseline with the Mann-Whitney odds as the estimand. The estimates and CIs can be calculated using the R package sanon (Kawaguchi and Koch 2015).

Additionally, the difference between treatment groups in the change from baseline at Week 3 and, separately, at delivery, will be evaluated using an ANCOVA model with fixed, categorical effects of treatment group, visit, and treatment group-by-visit interaction as well as the continuous, fixed covariates of baseline and the baseline-by-visit interaction. LS means will be calculated for each treatment group in the model. The difference between volixibat and placebo change from baseline in each outcome variable will be estimated, with the corresponding 2-sided 95% CI constructed for each visit. Change from baseline LS means with standard error, 95% CI for the LS means, p-value for testing if the LS mean is zero, LS mean difference between treatment groups (volixibat minus placebo) with standard error, 95% CI for the LS mean difference, and p-value for testing if the treatment LS means are equal will be calculated for each visit. The number of observations, mean, 95% CI on the mean, standard deviation, median, minimum, and maximum on both the observed and change from baseline values will also be summarized by treatment group for each visit.

For PGIC-Itch, since the PRO is a measure of change without a defined baseline, an ANCOVA model will be used, but will not include fixed covariates of baseline and the baseline-by-visit interaction. The dependent variable will be the PGIC-Itch value.

9. PHARMACOKINETIC ANALYSES

Due to poor absorption of volixibat very low systemic exposure and plasma drug levels are expected. Volixibat plasma concentrations will be summarized using descriptive statistics by analysis visit and dose of volixibat received at the specified analysis visit.

For maternal volixibat plasma concentration assessments: Plasma concentration of volixibat near predicted maximum observed concentration (C_{max}) will be obtained during the Week 1 visit, approximately 3 hours after dosing. Trough plasma levels of volixibat will be obtained during the Week 3 visit before dosing. Contingent upon meaningful exposures of volixibat, PK endpoints may include but are not limited to C_{max}, time to reach C_{max} (T_{max}), and the area under the plasma concentration-time curve (AUC) from time zero to the last measurable time point. Actual dosing and sampling times will be used for calculation of PK parameters. Additional PK parameters may be calculated.

For fetal (umbilical cord) volixibat plasma concentration assessments: Plasma concentration of volixibat in the umbilical cord blood sample obtained at delivery will be collected (when possible) and analyzed.

Plasma volixibat concentrations will also be presented in participants listings.

10. REFERENCES

Chan ISF, Zhang Z. Test-based exact confidence intervals for the difference of two binomial proportions. Biometrics 1999;55:1202–9.

Cnaan A, Laird NM, Slasor P. Using the general linear mixed model to analyse unbalanced repeated measures and longitudinal data. Stat Med 1997;16:2349–80.

European Medicines Agency. Guideline on missing data in confirmatory clinical trials. 2010.

Ge M, Durham LK, Meyer RD, et al. Covariate-adjusted difference in proportions from clinical trials using logistic regression and weighted risk differences. Drug Inf J 2011;45:481–93.

Kawaguchi A, Koch GG. Sanon: an R package for stratified analysis with nonparametric covariable adjustment. J Stat Softw 2015;671–37.

Kenward MG, Roger JH. Small sample inference for fixed effects from restricted maximum likelihood. Biometrics 1997;53:983–97.

Kieser M, Friede T. Simple procedures for blinded sample size adjustment that do not affect the type I error rate. Stat Med 2003;22:3571-81.

Koch GG, Carr GJ, Amara IA, Stokes ME, Uryniak TJ. Categorical Data Analysis. Chapter 13 in Berry, D.A. (ed.). Statistical Methodology in the Pharmaceutical Sciences. New York: Marcel Dekker, Inc., 1989:pp. 414-21.

Little R, Rubin D. Statistical analysis with missing data. New York: John Wiley. 1987.

Little R, Yau L. Intent-to-treat analysis for longitudinal studies with drop-outs. Biometrics 1996;52:1324–33.

Mallinckrodt CH, Clark WS, David SR. 2001. Accounting for dropout bias using mixed-effects models. J Biopharm Stat 2001*;*11:9–21.

Mallinckrodt CH, Lane PW, Schnell D, Peng Y, Mancuso JP. Recommendations for the Primary Analysis of Continuous Endpoints in Longitudinal Clinical Trials. Drug Information Journal 2008;42:303–19.

Mallinckrodt CH, Roger J, Chuang-Stein C, Molenberghs G, O'Kelly M, Ratitch B, Janssens M, Bunouf P. Recent developments in the prevention and treatment of missing data. Drug Information Journal 2013;48:68–80.

Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. The Annals of Mathematical Statistics 1947 Mar 1:50-60.

Molenberghs G, Kenward MG. Missing data in clinical studies*.* New York: Wiley. 2007.

Molenberghs G, Thijs H, Jansen I, et al. Analyzing incomplete longitudinal clinical trial data. Biostatistics 2004;5:445–64.

Pocock SJ, Simon R. Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. Biometrics 1975;31:103–15.

Qu Y, Zhao YD, Rahardja D (2008) Wilcoxon-Mann-Whitney test: stratify or not? J Biopharm Stat 18:1103-11.

Ratitch B, O'Kelly M. Implementation of pattern-mixture models using standard SAS/STAT procedures. Proceedings of PharmaSUG 2011 (Pharmaceutical Industry SAS Users Group), SP04, Nashville.

Ratitch B, O'Kelly M, Tosiello R. Missing data in clinical trials: from clinical assumptions to statistical analysis using pattern mixture models. Pharm Stat 2013;12:337–47.

Schafer JL. Analysis of Incomplete Multivariate Data. New Yori: Chapman & Hall. 1997.

Siddiqui O, Hung HM, O'Neill R. MMRM vs. LOCF: a comprehensive comparison based on simulation study and 25 NDA datasets. J Biopharm Stat 2009;19(2):227–46.

Stokes ME, Davis CS, Koch GG. Categorical data analysis using SAS. SAS Institute. 2012 Jul 31.

van Elteren PH. On the combination of independent two sample tests of Wilcoxon. Bulletin of the International Statistical Institute 1960; 37:351–61.

Verbeke G, Molenberghs G. Linear Mixed Models for Longitudinal Data. New York: Springer. 2000.

Wilcoxon F. Individual comparisons by ranking methods. Biometrics Bulletin, 1, 80-83. Reprinted in the Bobbs-Merrill Reprint Series in the Social Sciences, S-541. 1945.

Yan X, Su XG. Stratified Wilson and Newcombe confidence intervals for multiple binomial proportions. Statistics in Biopharmaceutical Research 2010;2(3):329-35.

Zink RC, Koch GG. NParCov3: a SAS/IML macro for nonparametric randomization-based analysis of covariance. J Stat Sotw 2012;50:1–7.

11. SOFTWARE

SAS® Software Version 9.4. SAS Institute Inc., Cary, NC, USA.

R Software Version 4.0.3 (or later). The R Foundation. Vienna, Austria.

12. SAP REVISION

Revision Date (DD MMM YYYY)	Section	Summary of Revision	Reason for Revision

13. APPENDIX 1 SCHEDULE OF ACTIVITIES

Table 7 Schedule of Activities

Visit Name	Screening	Treatment Period				Follow-Up		Notes
		Baseline	Week 1, 2, 3, etc.	Early Term*	EOT/ Delivery	F1	F2 (EOS)	*Participants who early terminate from study drug should still continue with weekly visits through delivery, F1, and F2.
Day/Week (Window)	Day −10 to −1	0*	Every 7 Days (±1 Day) Until Delivery		Day of Delivery (+2 Days)	Phone Call: 14 Days after Delivery (±2 Days)	Clinic Visit: 28 Days after Delivery (±3 Days)	Some visits may be conducted as a home health or remote visit at the discretion of the investigator upon approval by the medical monitor or designee and in accordance with local and regional regulations. *Regardless of date of screening visit, baseline visit must occur no later than 35 weeks and 0 days gestational age.
Administrative Procedures								
Informed consent	X							
Eligibility Procedures								
Inclusion/exclusion criteria	X	X						
Demography	X							
Prior/concomitant treatment	X	X	X	X	X	X	X	
Medical history	X	X						Includes menstrual and obstetric history
Estimated gestational age assessment	X							Confirmation of eligibility based upon estimated gestational age required before randomization: CRL from early ultrasound or head circumference on ultrasound if CRL is >84 mm.
Hepatitis and HIV screen, TSH	X							
Dosing Procedures								
Randomization		X						
Study treatment dispensing		X	X					As needed
Study treatment collection/reconciliation			X	X	X			As needed

Visit Name	Screening	Treatment Period				Follow-Up		Notes
		Baseline	Week 1, 2, 3, etc.	Early Term*	EOT/ Delivery	F1	F2 (EOS)	*Participants who early terminate from study drug should still continue with weekly visits through delivery, F1, and F2.
Day/Week (Window)	Day –10 to –1	0*	Every 7 Days (±1 Day) Until Delivery		Day of Delivery (+2 Days)	Phone Call: 14 Days after Delivery (±2 Days)	Clinic Visit: 28 Days after Delivery (±3 Days)	Some visits may be conducted as a home health or remote visit at the discretion of the investigator upon approval by the medical monitor or designee and in accordance with local and regional regulations. *Regardless of date of screening visit, baseline visit must occur no later than 35 weeks and 0 days gestational age.
Efficacy Procedures								
Education and compliance assessment of pruritus assessments	X	X	X					Education regarding importance of eDiary compliance at screening visit; reinforcement regarding eDiary compliance at study visits (as needed).
Once daily Adult ItchRO score assessment	X	X	X	X	X			Participant completes ItchRO at home on eDiary up to and including the day of delivery.
Pruritus/QoL PRO assessments (5D itch scale, CSS, EQ-5D-5L, PROMIS SF v1.0-Fatigue 7a, PROMIS SF v1.0-Sleep Disturbance, PIS-Itch)		X	X*	X	X**			Participant completes at clinic, except CSS, which the investigator completes at clinic. All PRO assessments to be completed by participant, including PIS-Itch and PGIC-Itch, should be collected prior to additional procedures at study visits. *At Week 3 visit only **For participants who terminate from study drug early, this does not need to be repeated.
PGIC-Itch			X*	X	X**			Participant completes at clinic. *At Week 3 visit only **For participants who terminate from study drug early, this does not need to be repeated.
Other Procedures								
Health utilization		X		X			X	
Safety Procedures								
Drug and alcohol screen	X							
Hematology panel	X		X	X	X		X	
Chemistry	X		X	X	X		X	

Visit Name	Screening	Treatment Period				Follow-Up		Notes
		Baseline	Week 1, 2, 3, etc.	Early Term*	EOT/ Delivery	F1	F2 (EOS)	*Participants who early terminate from study drug should still continue with weekly visits through delivery, F1, and F2.
Day/Week (Window)	Day –10 to –1	0*	Every 7 Days (±1 Day) Until Delivery		Day of Delivery (+2 Days)	Phone Call: 14 Days after Delivery (±2 Days)	Clinic Visit: 28 Days after Delivery (±3 Days)	Some visits may be conducted as a home health or remote visit at the discretion of the investigator upon approval by the medical monitor or designee and in accordance with local and regional regulations. *Regardless of date of screening visit, baseline visit must occur no later than 35 weeks and 0 days gestational age.
Physical examination	X	X	X	X	X		**X**	Complete examination conducted at screening and F2 visits. A symptom-directed examination focused on changes from last examination conducted at other visits. Includes confirmation of fetal cardiac activity via Doppler/auscultation at each visit while participant is pregnant.
Height	X							
Weight	X	X	X	X	X		X	
ECG	X							
Vital signs	X	X	X	X	X		X	
Coagulation (PT/INR, PTT)	X		X	X	X		X	
Vitamin A, D, and E levels		X	X*		X		X	*Every 4 weeks, starting with Week 4 visit. The sample collection tube should be protected from light.
Urinalysis	X	X	X	X	X		X	
Lipid panel, glycated albumin	X		X*		X		X	*Every 2 weeks, starting with Week 2 visit
sBA	X	X*	X*	X	X**		X	At each visit except EOT/Delivery, must be drawn in the non-fasting state (e.g., postprandially), 30-90 minutes after start of a meal. Every effort should be made to collect postprandial sBA levels at approximately the same time of day (morning, midday, or afternoon) at each visit starting with baseline, except for EOT/delivery. *At baseline and Week 3, an additional fasting sBA must be performed. **At EOT/Delivery, sBA may be performed without regard to last meal.
AE review		X	X	X	X	X	X	AEs for mother and baby to be collected and reviewed for up to 30 days after respective dates of hospital discharge.

Visit Name	Screening	Treatment Period				Follow-Up		Notes
		Baseline	Week 1, 2, 3, etc.	Early Term*	EOT/ Delivery	F1	F2 (EOS)	*Participants who early terminate from study drug should still continue with weekly visits through delivery, F1, and F2.
Day/Week (Window)	Day −10 to −1	0*	Every 7 Days (±1 Day) Until Delivery		Day of Delivery (+2 Days)	Phone Call: 14 Days after Delivery (±2 Days)	Clinic Visit: 28 Days after Delivery (±3 Days)	Some visits may be conducted as a home health or remote visit at the discretion of the investigator upon approval by the medical monitor or designee and in accordance with local and regional regulations. *Regardless of date of screening visit, baseline visit must occur no later than 35 weeks and 0 days gestational age.
Fetal ultrasound	X*	X*	X*					*Every 4 weeks while on study treatment; first ultrasound may be done during screening OR baseline (both are not required)
Antepartum fetal monitoring (NST/CTG or BPP)		X*	X*					*At baseline and Week 1 visits only, and thereafter at the discretion of the investigator
Perinatal outcomes assessment					X			Outcomes include • Gestational age at delivery • Mode of delivery (e.g., vaginal, operative vaginal, Cesarean, etc.; and iatrogenic vs. spontaneous) • Indication(s) for delivery, if iatrogenic • Apgar score at 5 minutes • Infant/placenta weight • Maternal estimated blood loss at delivery • Presence/absence of meconium staining of amniotic fluid • Neonatal diagnoses and data, including NICU data if applicable
Umbilical cord sample (at delivery, when possible)					X			Umbilical artery pH should be assessed first, followed by volixibat pharmacokinetics. Thereafter, sBA, total cholesterol, LDL cholesterol, and glucose can be assessed in any order.
Genetics - maternal blood (optional)		X						Specific testing TBD. Must confirm specific consent has been obtained for genetic testing
Genetics – cord blood sample (optional)					X			Specific testing TBD. Must confirm specific consent has been obtained for genetic testing.
NICU admission					X	X	X	Collect NICU admissions diagnoses and data for the duration of NICU stay (i.e., follow to resolution).

Visit Name	Screening	Treatment Period				Follow-Up		Notes
		Baseline	Week 1, 2, 3, etc.	Early Term*	EOT/ Delivery	F1	F2 (EOS)	*Participants who early terminate from study drug should still continue with weekly visits through delivery, F1, and F2.
Day/Week (Window)	Day −10 to −1	0*	Every 7 Days (±1 Day) Until Delivery		Day of Delivery (+2 Days)	Phone Call: 14 Days after Delivery (±2 Days)	Clinic Visit: 28 Days after Delivery (±3 Days)	Some visits may be conducted as a home health or remote visit at the discretion of the investigator upon approval by the medical monitor or designee and in accordance with local and regional regulations. *Regardless of date of screening visit, baseline visit must occur no later than 35 weeks and 0 days gestational age.
Drug Levels/Biomarkers Procedures								
Volixibat concentration sample (maternal plasma)			X*	X	X			*At Week 1 visit, 3 hours (+/- 15 minutes) post-dose; at Week 3 visit, ≤60 minutes pre-dose When feasible, maternal sample should be collected as close to time of actual delivery as possible. Drug level sampling may be discontinued in Part 2 pending Part 1 results
Volixibat concentration sample (fetal plasma from umbilical cord at delivery, when possible)					X			Drug level sampling may be discontinued in Part 2 pending Part 1 results
Biomarker samples (maternal) • C4 (fasting and postprandial, 30-90 min) • FGF-19 (fasting) • Autotaxin (fasting) • Progesterone sulfate • GLP-1(fasting)		X*	X*		X**			*Baseline and at Week 3 visit only **Random C4 only, timed with sBA draw at EOT/Delivery visit

5-D Itch; Adult Itch RO; AE=adverse event; BPP=biophysical profile; C4=7α-hydroxy-cholesten-3-one; CRL=crown rump length; CSS=Clinician Scratch Scale; CTG=cardiotocography; eDiary=electronic diary; EOS=end of study; EOT=end of treatment; EQ-5D-5L=EuroQoL 5-Dimension 5-Level; ET=early termination; F1=follow-up Visit 1; F2=follow-up Visit 2; ItchRO=Itch Reported Outcome; NICU=neonatal intensive care unit; NST=nonstress test; PGIC-Itch=Patient Global Impression of Change-Itch; PIS-Itch=Patient Impression of Severity of Itch; PRO=patient-reported outcome; PROMIS=Patient-Reported Outcomes Measurement Information System; PT=prothrombin time; PTT=partial thromboplastin time; QoL=quality of life; sBA=serum bile acids; SF=short form; TBD=to be determined; TSH=thyroid-stimulating hormone.

14. APPENDIX 2 ADVERSE EVENTS OF SPECIAL INTEREST AND CLINICAL INTEREST

The following event is an AESI for participants in this study:

- Symptomatic cholelithiasis
- Threatened preterm labor

The following events are ECIs in this study:

- Diarrhea (CTCAE Grade ≥3)
    - MedDRA preferred terms include the following:
        - Diarrhoea
- Abdominal pain (CTCAE Grade ≥3)
    - MedDRA preferred terms include the following:
        - Abdominal pain
        - Abdominal pain lower
        - Abdominal pain upper
- Transaminase increases (CTCAE Grade ≥3)
    - MedDRA preferred terms include the following:
        - Alanine aminotransferase abnormal
        - Alanine aminotransferase increased
        - Aspartate aminotransferase abnormal
        - Aspartate aminotransferase increased
        - Bilirubin conjugated abnormal
        - Bilirubin conjugated increased
        - Blood bilirubin abnormal
        - Blood bilirubin increased
        - Blood bilirubin unconjugated increased

        - Liver function test abnormal
        - Liver function test increased
        - Transaminases abnormal
        - Transaminases increased

- Fat-soluble vitamin deficiency
    - MedDRA preferred terms include the following:
        - Vitamin A decreased
        - Vitamin A abnormal
        - Vitamin A deficiency
        - Vitamin D decreased
        - Vitamin D abnormal
        - Vitamin D deficiency
        - Vitamin E decreased
        - Vitamin E abnormal
        - Vitamin E deficiency
        - Vitamin K decreased
        - Vitamin K abnormal
        - Vitamin K deficiency
        - International normalised ratio increased
        - International normalised ratio abnormal

15. APPENDIX 3 SAS SAMPLE CODES

A sample of the SAS code for sensitivity analysis using the copy-reference MI method is provided as below:

```
/*Generate transpose of adcf - wide format before Step 1*/
/*---- Generates 10 imputed data sets based on original (non-transformed)
data----*/
/*Step 1: Achieve Monotone Missing Data Pattern */
 PROC MI DATA=&DATA SEED=14823 NIMPUTE=10 OUT=MONO;
 /* Use numeric TRT01PN */
 VAR BASE AVAL1 AVAL2 AVAL3; /*Refer to AVAL at Visits 5, 6, and 7*/
 MCMC IMPUTE=MONOTONE NBITER=5000 NITER =5000;
 BY TRT01PN;
 run;

/*Step 2: Achieve Control-Based Copy-Reference Imputation */
 PROC MI DATA=MONO SEED=14823 NIMPUTE=1 OUT=OUTM1;
 CLASS TRT01PN;
 By _IMPUTATION_;
 VAR TRT01PN BASE AVAL1 AVAL2 AVAL3;
 MONOTONE REG (AVAL2 AVAL3);
 MNAR MODEL (AVAL1 AVAL2 AVAL3/ MODELOBS = (TRT01PN = '0'));
 /*Control=Placebo*/
 RUN;
 PROC SORT DATA =OUTM1; BY _IMPUTATION_ TRT01PN USUBJID LBASE; RUN;
 /*Generate transpose of OUTM1 - skinny format before Step 3*/

/*Step 3: Run MMRM Analysis on Imputed Data */
PROC MIXED DATA=OUTM1 METHOD=REML;
 BY _IMPUTATION_;
CLASS TRT01PN USUBJID AVISITN COUNTRY SEX;
MODEL CHG = TRT01PN AVISITN COUNTRY SEX TRT01PN*AVISITN BASE / DDFM=KR;
REPEATED / SUBJECT=USUBJID (TRT01PN) TYPE=UN;
LSMEANS TRT01PN*AVISITN / CL DIFF PDIFF;
 ESTIMATE 'TRT S - P AT WEEK 24' TRT01PN -1 1
 TRT01PN*AVISITN 0 0 -1 0 0 1 / CL LOWER LOWERTAILED;
 ODS OUTPUT PARAMETERESTIMATES=PESTIMATES;
RUN;

/*Step 4: Combining estimates from each imputed data set */
/*First, sort the data of estimates by visit (variable "LABEL") */
PROC MIANALYZE DATA=PESTIMATES;
 MODELEFFECTS ESTIMATE;
 STDERR STDERR;
 BY LABEL;
 ODS OUTPUT PARAMETERESTIMATES= MIPARM;
RUN;
```

A sample of the SAS code for sensitivity analysis using the tipping-point approach is provided as below:

```
/*-----------------------------------------------------------*/
```

```
/*--- Performs multiple imputation analysis ---*/
/*--- for specified shift parameters: ---*/
/*--- data= input data set ---*/
/*--- smin= min shift parameter ---*/
/*--- smax= max shift parameter ---*/
/*--- sinc= increment of the shift parameter ---*/
/*--- outparms= output parameters ---*/
/*-----------------------------------------------------------*/
%MACRO MIPARMS( DATA=, SMIN=, SMAX=, SINC=, OUTPARMS=);
DATA &OUTPARMS;
 SET _NULL_;
RUN;
/*------------ # of shift values ------------*/
%LET NCASE= %SYSEVALF( (&SMAX-&SMIN)/&SINC, CEIL );
/*---- Multiple imputation analysis for each shift ----*/
%DO JC=0 %TO &NCASE;
 %LET SJ= %SYSEVALF( &SMIN + &JC * &SINC);

 /*---- Generates 10 imputed data sets ----*/

/*Step 1: Achieve Monotone Missing Data Pattern */

 PROC MI DATA=&DATA SEED=14823 NIMPUTE=10 OUT=MONO;
 ** Use numeric TRT01PN ;
 VAR BASE AVAL1 AVAL2 AVAL3; /*Refer to AVAL at Visits 5, 6, and 7*/
 MCMC IMPUTE=MONOTONE NBITER=5000 NITER =5000;
 BY TRT01PN;
 run;

/*Step 2: Impute all visits under MAR first, then apply delta adjustments
at each visit*/

 PROC MI DATA= OUTM1 SEED=14823 NIMPUTE=1 OUT=OUTM2;
 CLASS TRT01PN;
 BY _IMPUTATION_;
 VAR TRT01PN BASE AVAL1 AVAL2 AVAL3;
 MONOTONE REG (AVAL2 AVAL3);
 MNAR ADJUST( AVAL1 / SHIFT=&SJ ADJUSTOBS=(TRT01PN ='1') );
 MNAR ADJUST( AVAL2 / SHIFT=&SJ ADJUSTOBS=(TRT01PN ='1') );
 MNAR ADJUST( AVAL3 / SHIFT=&SJ ADJUSTOBS=(TRT01PN ='1') );
 RUN;
 PROC SORT DATA =OUTM1; BY _IMPUTATION_ TRT01PN USUBJID BASE; RUN;

/*Generate transpose of OUTM2 - skinny format before Step 3*/

/*Step 3: Run MMRM Analysis on Imputed Data */

PROC MIXED DATA=OUTM2 METHOD=REML;
 BY _IMPUTATION_;
CLASS TRT01PN USUBJID AVISITN COUNTRY SEX;
MODEL CHG = TRT01PN AVISITN COUNTRY SEX TRT01PN*AVISITN BASE /
DDFM=KR;
REPEATED / SUBJECT=USUBJID (TRT01PN) TYPE=UN;
LSMEANS TRT01PN*AVISITN / CL DIFF PDIFF;
  ESTIMATE 'TRT S - P AT WEEK 24' TRT01PN -1 1
 TRT01PN*AVISITN 0 0 -1 0 0 1 / CL LOWER LOWERTAILED;
 ODS OUTPUT PARAMETERESTIMATES=PESTIMATES;
```

```
 RUN;

 /*First, sort the data of estimates by visit (variable "LABEL") */

 PROC SORT DATA=PESTIMATES; BY LABEL; RUN:

 /*Step 4: Combining estimates from each imputed data set */

PROC MIANALYZE DATA=PESTIMATES;
 MODELEFFECTS ESTIMATE;
 STDERR STDERR;
 BY LABEL;
 ODS OUTPUT PARAMETERESTIMATES=MIPARM;
RUN;
DATA MIPARM;
SET MIPARM;
 SHIFT= &SJ;
RUN;

/*----- Output multiple imputation results ----*/

DATA &OUTPARMS;
 SET &OUTPARMS MIPARM;
RUN;
%END;
%MEND MIPARMS;

/*Assume that the tipping point for the shift parameter that reverses the
study conclusion is between 0 and 100. The following statement performs
multiple imputation analysis for each of the shift parameters
0,10,20,30,…,100.*/

ODS LISTING CLOSE;
%MIPARMS(DATA=XXXX, SMIN=0, SMAX=100, SINC=10, OUTPARMS=PARMS1);

/* Step 5: Finding Tipping Point for Shift Parameter between 0 and 100
such that p-value is less than 0.05, below is an example for a tipping
point for the shift parameter between 40 and 50.
For a two-sided Type I error level of 0.05, the tipping point for the
shift parameter is greater than 40. The following statement performs
multiple imputation analysis for shift parameters 40, 41, …, 50*/

%MIPARMS(DATA=XXXX, SMIN=40, SMAX=50, SINC=1, OUTPARMS=PARM2);
/*The following statements display the p-values that are associated with
the shift parameters.*/
PROC PRINT LABEL DATA=PARM2;
 VAR SHIFT PROBT;
 TITLE 'P-VALUES FOR SHIFT PARAMETERS';
 LABEL PROBT='PR > |T|';
 FORMAT PROBT 8.4;
RUN;
```

A sample of the SAS code for van Elteren test is provided as below:

```
proc npar1way data=adeff ;
  strata BLGSTAGE BLITCHRO BLGDIAB;
  class trt01p;
  var chg;
run;
```