Statistical Analysis Plan

Protocol title:	**A Phase 3 Double-blind, Randomized, Placebo-controlled Study of the Safety and Efficacy of Odevixibat (A4250) in Patients with Alagille Syndrome (ASSERT)**
Protocol number:	**A4250-012**
Protocol Version/Date:	**V2.0 / 28August 2020**
IND Number:	**145988**
Study phase:	**Phase 3**
Date of issue:	**27 Sep 2022**
EudraCT number	**2020-004011-28**

Signature Page

PPD

27-Sep-2022 | 8:09 PM CDT

Statistician, Biostatistics
Firma Clinical Research

[Date]

PPD

27-Sep-2022 | 7:12 PM CDT

Senior Director, Biostatistics
Albireo Pharma, Inc

[Date]

PPD

27-Sep-2022 | 7:06 PM CDT

Medical Director
Albireo Pharma, Inc

[Date]

Table of Contents

List of Tables

List of Abbreviations

The following abbreviations will be used within this SAP

Abbreviation or Special Terms	Explanation
AE(s)	Adverse Event(S)
AFP	Alfa-fetoprotein
AEIs	Adverse Events of Interest
ALGS	Alagille Syndrome
ALT	Alanine Aminotransferase
ALP	Alkaline Phosphatase
ANCOVA	Analysis of Covariance
AR(1)	First-Order Autoregressive
AST	Aspartate Aminotransferase
ATC	Anatomical Therapeutic Chemical
BMI	Body Mass Index
CaGIC	Caregiver Global Impression of Change
CaGIS	Caregiver Gobal Impression of Symptom
CDC	Centers for Disease Control and Prevention
CGIC	Clinical Global Impression of Change
CGIS	Clinical Global Impression of Symptom
CMH	Cochran-Mantel-Haenszel
COVID-19	Coronavirus Disease 2019
CRO	Contract Research Organization
CS	Compound Symmetry
CSR	Clinical Study Report
DILI	Drug-Induced Liver Injury
DSMB	Data and Safety Monitoring Board
EAIR	Exposure Adjusted Incidence Rate
eCRF	Electronic Case Report Form
eDiary	Electronic Diary
eGFR	Estimated Glomerular Filtration Rate
EOT	End of Treatment
EOS	End of Study
EU	European Region
FAS	Full Analysis Set
FDA	United States Food and Drug Administration
FSV	Fat-Soluble Vitamin
GIC	Global Impression of Change

Abbreviation or Special Terms	Explanation
GIS	Global Impression of Symptoms
GGT	Gamma-Glutamyl Transferase
HTN	Hypertension
HSAC	Hepatic Safety Adjudication Committee
ICE	Intercurrent Event
ICF	Informed Consent Form
ICH	International Council for Harmonisation
IDMS	Isotope Dilution Mass Spectrometry
INR	International Normalized Ratio
IWRS	Interactive Web Response System
LFT	Liver Function Test
MAR	Missing at Random
MCMC	Markov Chain Monte Carlo
MedDRA	Medical Dictionary for Regulatory Activities
MDRD	Modification of Diet in Renal Disease
MH	Medical History
MI	Multiple Imputation
MMRM	Mixed-Effect Model for Repeated Measures
MNAR	Missing Not at Random
ObsRO	Observer Reported Outcome
p-C4	Plasma 7α-Hydroxy-4-Cholesten-3-One Concentration
PedsQL	Pediatric Quality of Life Inventory
PFIC	Progressive Familial Intrahepatic Cholestasis
PGIC	Patient Global Impression of Change
PGIS	Patient Global Impression of Symptoms
PK	Pharmacokinetics
PP	Per Protocol
PRO	Patient Reported Outcome
PT	Preferred Term
RoW	Rest of World
SAE(s)	Serious Adverse Event(s)
SAP	Statistical Analysis Plan
SAF	Safety Analysis Set
SAS	Statistical Analysis System
SD	Standard Deviation
SI	International System of Unit
SMQs	Standardised MedDRA Queries

Abbreviation or Special Terms	Explanation
SOC	System Organ Class
SSR	Sample Size Re-estimation
TEAE(s)	Treatment Emergent Adverse Event(s)
TFLs	Tables, Figures and Listings
UDCA	Ursodeoxycholic Acid
ULN	Upper Limit of Normal
UN	Unstructured Covariance Matrix
US	United States
WHO	World Health Organization
WHO-DD	World Health Organization-Drug Dictionary

Version History

This statistical analysis plan (SAP) for Study A4250-012 is based on protocol version 2.0 dated 28-Aug-2020. This section summarizes major changes to the statistical analysis features in the SAP. This SAP will be approved before the database lock.

Table 1 - Major changes in statistical analysis plan

SAP Version	Approval Date	Change	Rationale
1.0	14Oct2021	NA	Original version
2.0	27Sep2022	Section 2.1, Section 3 table 5, Section 4.1, Section 4.2, Section 4.7 (in SAP version 1.0) and Section 4.8: Remove analyses for exploratory cohort (≥18 years old)	There were no subjects enrolled in the exploratory cohort.
		Section 1.3 Table 3 and Section 4.5.2: Update primary estimand and secondary estimand. Keep one estimand with two strategies for addressing different intercurrent events.	Incorporate the comments from the FDA's Meeting Request – Written Responses question 2(a) - 2(c), dated 09MAR2022.
		Section 1.3 Table 3 and Section 4.5.2: remove "Change in anti-pruritus medication exceeding a specified duration or dose" as an ICE event and add rationale in section 4.5.2.	Incorporate the comments from the FDA's Meeting Request – Written Responses question 2(d), dated 09MAR2022.
		Section 1.3 Table 3 and Section 4.5.2.2.5: Add a sensitivity analysis for the primary endpoint utilizing the 28-day baseline data for the patients for whom it is available.	Incorporate the comments from the FDA's Meeting Request – Written Responses question 2(e), dated 09MAR2022.
		Section 1.3 Table 3 and Section 4.5.2.2.6: Add a sensitivity analysis evaluating scratching at Month 6 (Weeks 21 to 24) adjusting for baseline scratching in the analysis.	Incorporate the comments from the FDA's Meeting Request – Written Responses question 2(f), dated 09MAR2022.
		Section 4.5.2.2.1: For the multiple imputation (MI) sensitivity analysis assuming data is missing not at random (MNAR), update to Control-based MI by imputing missing data after all types of treatment discontinuations based on the outcomes of the placebo patients.	Incorporate the comments from the FDA's Meeting Request – Written Response to Question 2(g), dated 09MAR2022.
		Section 4.5.2.2.1 and Section 5.6: Provide details and the sample SAS codes for control-based imputation including "Only imputed values within the range of 0 to 4 will be used."	Incorporate the comments from the FDA's Meeting Request – Written Response to Question 2(h), dated 09MAR2022.
		Section 4.5.2.2.2: Add details on the tipping point analysis.	Incorporate the comments from the FDA's Meeting Request – Written Response to Question 2(h), dated 09MAR2022.

SAP Version	Approval Date	Change	Rationale
		Section 4.5.2.1 and Section 5.6: Add a supportive analysis and a sample SAS code to estimate the average treatment effect at Month 4 to 6 using the same mixed-effect model for repeated measures (MMRM) that is proposed for the current primary endpoint.	Incorporate the comments from the FDA's Meeting Request – Written Response to Question 4, dated 09MAR2022.
		Age calculation in Table 9:	Clarification on the calculation of age due to the data collection for site CCI .
		Section 4.5.3 and Section 4.6.3: Remove the age stratification and baseline direct bilirubin from the modeling for subgroup.	Due to the small sample size of the study.
		Table 5 and section of protocol deviations.	Clarification on the definitions of per protocol analysis set (PPS) and important protocol deviation.
		Section 4.4: update the formula for average daily dose	Correct the average daily dose (ug/kg/day) by dividing the weight in the formula.
		Section 4.4: modify the formula for treatment compliance rate.	Considering the dose interruption as non-compliance
		Worst weekly score in Table 9: adding baseline definition as the first dose date to 14 days prior to the first dose date and this is the baseline definition for the analysis proposed in the SAP.	To be consistent with the baseline definition for monthly scratching score.
		Section 4.6.4.2: add clinically meaningful threshold for pruritus responder analysis.	Results from blinded psychometric analysis.
		Table 6: update definition of AEIs	Incorporate the comments from the FDA's Meeting Request – 'Additional Comments', dated 09MAR2022.

1 Introduction

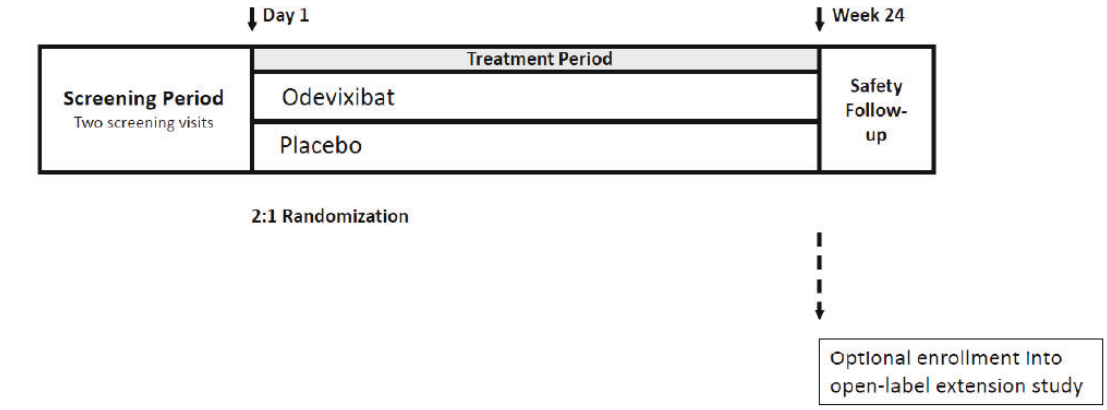
The purpose of this SAP is to provide detailed descriptions of the statistical methods, data derivations, and data displays for study protocol A4250-012 "A Phase 3 Double-blind, Randomized, Placebo-controlled Study of the Safety and Efficacy of Odevixibat (A4250) in Patients with Alagille Syndrome (ASSERT)". The table of contents and templates for the Tables, Figures and Listings (TFLs) will be produced in a separate document.

Any deviations from this SAP will be described and justified in the Clinical Study Report (CSR). The preparation of this SAP has been based on ICH E3, E9 and E9(R1) guidelines [1, 2, 3]. All data analyses and generation of TFLs will be performed using SAS® version 9.4 or higher. The SAP will be finalized and signed off prior to locking the database and unblinding.

1.1 Study Design

This is a phase 3, double-blind, randomized, placebo-controlled study to investigate the efficacy and safety of 120 µg/kg/day odevixibat compared to placebo in patients with Alagille syndrome (ALGS).

Figure 1 – Study Design



Odevixibat = once daily treatment with 120 µg/kg/day odevixibat

The study includes a screening period of up to 56 days, followed by a 24-week treatment period, and a Safety Follow-up Period. Eligible patients will be randomized 2:1 to receive odevixibat 120 µg/kg/day or placebo. An Interactive Web Response System (IWRS) will be used to assign patients to treatment. The randomization codes were computer generated by a qualified randomization vendor. An experimental cohort of approximately 45 patients <18 years of age at randomization with a genetically confirmed diagnosis of ALGS will be enrolled. Patients <18 years of age will be randomized according to an age stratification factor, i.e., <10, and 10 to <18 years of age. The protocol included an additional exploratory cohort of patients ≥18 years of age with a genetically confirmed diagnosis of ALGS, not to exceed 18 patients in total, who were also to be randomized 2:1 favoring active treatment. A separate randomization scheme was provided for this cohort. There were no subjects enrolled in this cohort at the time enrollment completed for the study.

Patients who complete the Week 24/EOT visit in this study will be invited to participate in an open-label extension study in which all patients will receive active treatment. Patients who enroll in the open-label extension study will not participate in the Safety Follow-up Period.

There will be two treatment groups in this study, odevixibat 120 µg/kg/day and placebo. Odevixibat and placebo will be supplied as capsules for oral administration. The number of capsules provided to the patient will be based on body weight at randomization (Study Day 1) and can be adjusted depending on the patient's weight (Protocol Table 2) at each study visit.

1.2 Objective and Endpoints

The primary, key secondary, and secondary efficacy objectives, and safety objectives and endpoints are listed in Table 2.

Table 2 - Objectives and endpoints

	Objectives	Endpoints
Primary Efficacy		
	• To demonstrate the efficacy of repeated daily doses of 120 µg/kg/day odevixibat in relieving pruritus in patients with ALGS	• Change from baseline in scratching score to Month 6 (Weeks 21 to 24) as measured by the Albireo Observer reported outcome (ObsRO) caregiver instrument
Key Secondary Efficacy		
	• To assess the impact of odevixibat on serum bile acid levels in patients with ALGS	• Change in serum bile acid levels from baseline to the average of Week 20 and Week 24
Secondary Efficacy		
		• Change from baseline in pruritus score to Month 6 (Weeks 21 to 24) as measured by the Albireo Patient reported outcome (PRO) instrument • Percentage of patients achieving a clinically meaningful decrease in pruritus (pruritus responders) as measured by the Albireo ObsRO/PRO instruments • Change from baseline through Week 24 in patient reported and observer reported itching and scratching severity scores, respectively, for the morning assessment and for the evening assessment. These endpoints will be assessed combining age groups, and by age group, 0 to <8, 8 to <12, and 12 to <18 • Change from baseline to Week 24 in sleep parameters as measured with the Albireo ObsRO/PRO instruments (e.g., tiredness and number of awakenings)

Objectives	Endpoints
	• Change from baseline to Week 24 in Pediatric Quality of Life Inventory (PedsQL) subdomain scores • Assessment of Global Symptom Relief from baseline to Weeks 4, 12, and 24 as measured by patient, caregiver, and clinician Global Impression of Symptoms (GIS) (PGIS, CaGIS, CGIS) items • Assessment of Global Symptom Relief as measured by patient, caregiver, and clinician Global Impression of Change (GIC) (PGIC, CaGIC, CGIC) items at Weeks 4, 12, and 24 • Patient impression of treatment effect as recorded during exit interviews at Week 24 • Change from baseline to Week 24 in xanthomatosis as assessed by the Clinician Xanthoma Scale • Change in serum bile acid levels from baseline through Week 24 • Change from baseline to Week 24 in serum alanine aminotransferase (ALT) concentration • Change from baseline to Week 24 in serum aspartate aminotransferase (AST) concentration • Change from baseline to Week 24 in gamma-glutamyl transferase concentration • Change from baseline to Week 24 in total bilirubin concentration • Change from baseline in biochemical markers and measures of bile acid synthesis (autotaxin, Plasma 7α-hydroxy-4-cholesten-3-one concentration (p-C4)) • Change from baseline in total cholesterol concentration
Safety	
• To evaluate the safety and tolerability of odevixibat in patients with ALGS	• Occurrence of treatment emergent adverse events (TEAEs) including severity and relatedness to study drug at all visits • The incidence of treatment emergent serious adverse events (SAEs), based upon information from patient reports, including the description, incidence, and severity of an SAE • Trends evaluated for the following assessments: physical examinations, concomitant medications, vital signs, laboratory test results (including clinical chemistry, hematology, urinalysis, alfa-fetoprotein, vitamins A and E, 25-hydroxy

Objectives	Endpoints
	vitamin D and International Normalized Ratio [INR])

1.3 Estimands

Estimand attributes and analyses for the primary efficacy endpoint and the key secondary efficacy endpoint are summarized in Table 3. More details are provided in Section 4.

Table 3 - Summary of estimands for the primary and the key secondary efficacy endpoints

ENDPOINT CATEGORY (ESTIMAND)	ESTIMAND					ANALYSIS AND MSSING DATA HANDLING
	VARIABLE	POPULATION	POPULATION-LEVEL SUMMARY	INTERCURRENT EVENT (ICE)	STRATEGY FOR ADDRESSING ICE	
Primary objective: To demonstrate the efficacy of repeated daily doses of 120 µg/kg/day odevixibat in relieving pruritus in patients <18 years of with ALGS.						
Primary efficacy endpoint (Estimand)	Change from baseline in average AM and PM scratching score to Month 6 (Weeks 21 to 24)	FAS	The difference between treatment with odevixibat and placebo in mean change in scratching score from baseline to Month 6 (Weeks 21 to 24)	Premature discontinuation of study drug treatment prior to Week 24	Treatment Policy (include data after ICE)	Main analysis (PRIMARY ANALYSIS): A MMRM modelling change from baseline for each four-week average AM and PM scratching score, with baseline age stratification, baseline average AM and PM scratching score, baseline direct bilirubin, treatment group, time (in months), and treatment-by-time interaction in the model. No imputation is implemented. (section 4.5.2.1) Sensitivity analysis 1: MMRM with control-based multiple imputation (MI) (section 4.5.2.2.1) Sensitivity analysis 2: Tipping point analysis (section 4.5.2.2.2) Sensitivity analysis 3: MMRM based on worst weekly scratching score for a month (section 4.5.2.2.3) Sensitivity analysis 4: MMRM based on worst weekly scratching score for a month with control-based MI (section 4.5.2.2.4) Sensitivity analysis 5: MMRM for the endpoint with the baseline pruritus score calculated based on 28 days prior to first dose date. (section 4.5.2.2.5) Sensitivity analysis 6: MMRM for the scratching score
				Undergoes biliary diversion surgery or liver transplant	Hypothetical (exclude data after ICE)	

ENDPOINT CATEGORY (ESTIMAND)	ESTIMAND					ANALYSIS AND MSSING DATA HANDLING
	VARIABLE	POPULATION	POPULATION-LEVEL SUMMARY	INTERCURRENT EVENT (ICE)	STRATEGY FOR ADDRESSING ICE	
						at Month 6 (Weeks 21 to 24) adjusting for baseline scratching in the analysis. (section 4.5.2.2.6) Supplementary analysis: The same MMRM utilized for the main analysis will also be used for the PP analysis set (section 4.5.2.3)
Key secondary objective: To assess the impact of odevixibat on serum bile acid levels in patients <18 years of age with ALGS.						
Key secondary efficacy endpoint: (Estimand)	Change in serum bile acid levels from baseline to the average of Week 20 and Week 24	FAS	The difference between treatment with odevixibat and placebo in mean change in serum bile acid levels from baseline to the average of Week 20 and Week 24	Premature discontinuation of study drug treatment prior to Week 24	Treatment Policy	Main analysis (PRIMARY ANALYSIS): The MMRM utilized for the primary efficacy endpoint will also be used for assessment of the key secondary endpoint (section 4.6.2.1) Sensitivity analysis 1: MMRM with control-based MI (section 4.6.2.2.1) Sensitivity analysis 2: Tipping point analysis (section 4.6.2.2.2) Supplementary analysis 1: Rank transform analysis of covariance (ANCOVA) (section 4.6.2.3.1) Supplementary analysis 2: The same MMRM utilized for the main analysis will also be used for the PP analysis set (section 4.6.2.3.2)
				Undergoes biliary diversion surgery liver transplant	Hypothetical	

2 Sample Size Determination

2.1 Sample Size and Power

Forty-five (45) patients <18 years of age were randomized at an experimental-to-control allocation of 2 to 1 in order to obtain approximately 36 completers, assuming an approximate drop-out rate of 20%. At a 1-sided significance level of $alpha_{1\text{-sided}} = 0.025$, assuming a pooled standard deviation (SD) of 1.0, and a difference between the treatment groups of 1.2 in mean change in pruritus score from baseline to Month 6 favoring the experimental arm, the power of the study is 0.909, using the exact method (Proc Power, SAS v.9.4, Cary, NC). The key secondary endpoint is also powered for a standardized treatment effect (treatment effect/SD) of 1.2.

2.2 Blinded Sample Size Re-Estimation

Blinded sample size re-estimation (SSR) is classified as a well-understood adaptive feature in the FDA guidance [4] because it does not involve treatment unblinding and can be conducted without worrying about introducing bias into a trial. This one-sample variance SSR method [5] is based on non-comparative blinded data for which no alpha adjustment is required. If the pooled SD was overestimated in the original sample size calculation, the sample size will not be decreased. If the pooled SD was underestimated in the original sample size calculation, the sample size will increase up to 54 to maintain the study power ≥0.80 to detect a standard effect size of ≥0.83 at a significance level of $alpha_{1\text{-sided}} = 0.025$, assuming the same 1.2 treatment difference in mean change in pruritus score from baseline to Month 6 (Week 21 to 24) as the original sample size calculation.

After a minimum of 18 patients in the experimental cohort have completed the Week 16 visit, the pooled SD of change from baseline to Month 4 (Weeks 13 to 16 assessments) and available change from baseline to Month 6 (Week 21 to 24) in average monthly scratching score will be calculated by an independent blinded statistician from Contract Research Organization (CRO), Firma Clinical. The independent blinded statistician will provide the following statement to the Sponsor: "The observed pooled SD at Week 16 and Week 24 available data is x.xxx. This estimate is based on xx subjects with non-missing outcome at Week 13 to 16 and xx subjects with non-missing outcomes at Week 21 to 24." The process of the blinded SSR will be documented by Firma Clinical in the Blinding and Unblinding Plan for Protocol A4250-012 prior to performing the blinded SSR.

CCI patients are expected to complete Visit Week 24 at the time of the blinded SSR. Based on scratching data from Study A4250-005 in the progressive familial intrahepatic cholestasis (PFIC) indication, the Month 6 pooled SD from the CCI patients is approximately CCI of the pooled SD based on Month 4 data from the CCI patients and Month 6 (Week 21 to 24) data from the CCI patients. Therefore, a multiplication factor CCI is applied when calculating the standard effect size at Week 24 in Table 4. The Sponsor will determine the number of completers per the sample size calculation listed in Table 4. The number of additional patients to be enrolled will be based on the actual dropout rate.

Table 4 – Blinded sample size re-estimation

Observed Pooled SD at Week 16 & Week 24 available data	Standard effect size at Week 24	N	Power
≤0.95	≥1.18	36	≥0.90
>0.95 to ≤0.98	≥1.14 to <1.18	39	≥0.90
>0.98 to ≤1.03	≥1.09 to <1.14	42	≥0.90
>1.03 to ≤1.10	≥1.02 to <1.09	45	≥0.88 to <0.90
>1.10 to ≤1.17	≥0.96 to <1.02	48	≥0.86 to <0.88
>1.17 to ≤1.25	≥0.90 to <0.96	51	≥0.84 to <0.86
>1.25 to ≤1.35	≥0.83 to <0.90	54	≥0.80 to <0.84
>1.35	<0.83	54	<0.80

Assuming a significance level of alpha1-sided = 0.025 and 1.2 treatment difference in mean change in scratching score from baseline to Month 6
N: Number of completers.

3 Analysis Populations

The following populations for analyses are defined:

Table 5 - Populations for analyses

Population	Description
Randomized	All patients from the screened population who have been allocated to a randomized intervention by IWRS regardless of whether the intervention was received or not.
Full Analysis Set (FAS)	All randomized patients who received at least 1 dose of study drug treatment. Patients will be analyzed in the treatment group they were randomized to even if they received the incorrect study drug. The FAS will be utilized for efficacy analyses.
Per Protocol Analysis Set (PPS)	The PPS is a subset of the FAS and will consist of all randomized patients for whom no important protocol deviation (i.e., no major protocol deviation which may affect the study efficacy outcomes) is documented. In addition, patients meet following criteria will be excluded from PPS. • Overall treatment compliance is less than 70% • Treatment duration is less than 50% of the days with eDiary data collected during weeks 21 to 24 The PPS will provide supportive data for the efficacy analyses. The algorithm for PPS will be defined and allocation of patients to the PPS will be performed before the database lock and un-blinding of the study.
Safety Analysis Set (SAF)	All patients who received at least 1 dose of study drug. Patients will be analyzed according to the treatment they actually received for the first dose. The SAF will be utilized for safety analyses.

A listing for each analysis set will be provided. The reason for exclusion from PPS will be summarized by treatment groups and will be included in a listing.

4 Statistical Analyses

4.1 General Considerations

All personnel involved with the analysis of the study and the independent blinded statistician who performs the blinded SSR will remain blinded until database lock except the unblinded Data Safety Monitoring Board (DSMB) statistician/programmers from Firma who are independent from both the Firma study team and Albireo. All statistical procedures will be completed using SAS® version 9.4.

In general, continuous data will be summarized using descriptive statistics, including the number of observations available (n), mean, SD, median, minimum, and maximum. The letter "n" will be presented without a decimal point. Minimum and maximum values will be presented to the same precision as in the database. Mean and median will be presented to 1 more decimal place than the minimum and maximum. SD will be presented to 1 more decimal place than the mean and median.

Categorical and ordinal data will be summarized using the count (frequencies) and percentage of patients. Percentages will be rounded to 1 decimal place. Descriptive summaries of change from baseline in categorical variables will be provided using shift tables, as applicable.

For summary purposes, if not otherwise specified, the baseline value of a parameter is defined as the last available assessment of that parameter prior to the first intake of study drug. Derived variables used for the analyses are provided in Appendix 2. All summaries will be presented by treatment group, unless otherwise specified.

For summary tables, if not otherwise specified, the following 2 presentations will be used:

- For baseline related tables and protocol deviations: by treatment group (odevixibat and placebo) and overall (2 treatment groups combined) will be presented.
- For the safety and efficacy related tables: by treatment group (odevixibat and placebo) will be presented.

The analysis window details for the derivation of study weeks are outlined in Appendix 3.

Age is the stratification factor for the study. Since the purpose of stratification in the randomization is to balance the treatments in each stratum and following the intend-to-treat principle, for any analysis based on the FAS where strata is used for covariate adjustment of the modeling (with the exception of subgroup analyses), the strata of age that is recorded in the randomization system (IWRS) will be used. For all other analyses based on the FAS and all subgroup analyses, the actual age will be used to categorize the age groups.

Selected patient data, including derived data, will be presented in individual patient data listings. All listings will be sorted by treatment group, patient number, date/time and visit. Each patient's sex and age at randomization will be stated on each listing. Data listings will be based on all randomized patients.

4.2 Patient Dispositions

The number (%) of patients in the following categories will be provided:

- Patients randomized
- Patients dosed
- Patients not dosed
- Patients completing treatment
- Patients discontinuing treatment early (including reason for discontinuation of treatment)
- Patients completing Study A4250-012
  - Patients discontinuing treatment
  - Patients enrolled into Study A4250-015
  - Patient did not enroll into Study A4250-015
- Patients withdrawing early from the study (including withdrawal reason)

Additionally, patients enrolled, included in the SAF, in the FAS, and in the PP analysis set will be summarized by region (Appendix 4).

Summary table for patients with premature discontinuation of study drug treatment, undergoing biliary diversion surgery and liver transplantation will be provided.

Protocol deviations

Important protocol deviations (IPD) will be identified by study team according to the criteria defined prior to database lock. IPDs will include following categories:

- Failure to meet following inclusion or exclusion criteria
  - Inclusion criteria (2) Patient must have a history of significant pruritus and a caregiver reported observed scratching or a patient -reported pruritus score at an average of ≥ 2
  - Inclusion criteria (3) Patient must have an elevated baseline serum bile acid level.
- Non-compliance in study procedure
  - Missing pruritus score at month 6 due to insufficient eDiary data collected during weeks 21-24

The number of patients with protocol deviations due to COVID-19 and number of protocol deviation due to COVID-19 will be summarized in the FAS.

In addition, a listing will be provided with all protocol deviations identified based on data recorded on protocol deviation logs (on randomized patients) prior to database lock.

4.3 Demographics and Baseline Characteristics, Prior or Concomitant Medications

Demographics, baseline and disease characteristics

The following demographics, baseline and disease characteristics, medical and surgical history will be summarized using descriptive statistics in the FAS.

Demographic, baseline, and disease characteristics

- Age in years as quantitative variable and Age in years as categorical variables (Age group 1: 0 to <10, 10 to <18; Age group 2: 0 to <2, 2 to <12, 12 to <18) based on actual age.
- Age group (0 to <10, 10 to <18) from IWRS for randomization
- Gender (Male, Female)
- Race (American Indian or Alaska Native, Asian, Black or African American, Native Hawaiian or Other Pacific, Islander, White, Other, Not reported)
- Race group (White, Non-White (including race as not reported))
- Ethnicity (Hispanic or Latino, Not Hispanic or Latino, Not reported)
- Ethnicity group (Hispanic or Latino (including ethnicity as not reported), Not Hispanic or Latino)
- Country and region (Appendix 4)
- Use of ursodeoxycholic acid (UDCA)
- Use of anti-pruritus medication
- Baseline values: height (or length depending on age), weight as quantitative and categorical (≤10 and >10 kg) variable, body mass index (BMI), Z-score (height, weight, and BMI), pruritus score (i.e., AM, PM, average AM and PM scratching score from ObsRO), serum bile acid as quantitative and categorical (≥ median and < median based on overall randomization population) variable, estimated glomerular filtration rate (eGFR), ALT, AST, gamma-glutamyl transferase (GGT), total bilirubin, direct bilirubin as quantitative and categorical (>3 and ≤3 mg/dL), alkaline phosphatase (ALP), INR, cholesterol, triglycerides, vitamin A, vitamin E, and 25-hydroxy vitamin D.
- Disease characteristics include
  - Genetic testing results for ALGS if performed at screening (e.g., mutation in *JAG1* vs *NOTCH2*; per protocol, testing is only performed at screening if not available in the patient's medical record), years since ALGS diagnosis
  - Child-Pugh classification (mild (Class A), moderate (Class B), or severe hepatic impairment (Class C)), and hepatic impairment classification (no impairment, mild, moderate, or severe) per NCI Organ Dysfunction Working Group [6]. Details of the classification are described in Appendix 9.

eGFR will be calculated based on the modified/bedside Schwartz equation. The values of eGFR will be provided by the central laboratory vendor.

Listings will be created to present all the demographic, baseline, and disease characteristics for all randomized patients.

Medical surgical history

Medical and surgical history includes date of diagnosis of ALGS, prior medications for ALGS, historical (2 years prior to Screening) liver function test (LFT) values (AST, ALT, and total bilirubin), previous genetic testing for ALGS, and historical biopsy data.

Medical and surgical history will be coded to a preferred term (PT) and associated primary System organ class (SOC) using the Medical Dictionary for Regulatory Activities (MedDRA) version 25.0. The frequencies and percentages of patients with reported medical and surgical history will be presented by SOC and PT for the SAF. The summary table will be sorted alphabetically for SOC and PT. Medical and surgical history will also be listed.

Prior or concomitant medications

All medications will be coded using the World Health Organization-Drug Dictionary (WHO-DD) using the version March 2022.

- Prior medications are any medications taken by the patient within 3 months prior to first intake of study drug that is recorded on the eCRF. Prior medications can be discontinued before the first intake of study drug or can be ongoing during the treatment period.
- Concomitant medications are any medications received by the patient concomitantly on or after the first intake of study drug.

Prior and concomitant medications will be summarized descriptively using frequency tables by Anatomical Therapeutic Chemical (ATC) and PT for the SAF. ATC class 4 and PT will be presented alphabetically. All prior and concomitant medications will be listed. Details for imputing missing or partial start and/or stop dates of non-study medications are described in Appendix 5.

4.4 Extent of Exposure

The extent of exposure will be assessed by the duration of treatment exposure, average daily dose, and compliance. Exposure will be summarized with descriptive statistics (n, mean, median, SD, minimum, and maximum) and presented using the SAF. In addition, duration of treatment exposure will be summarized categorically: <= 4 weeks, >4 – <= 8 weeks, >8 – <= 12 weeks, >12 – <= 16 weeks, >16 – <= 20 weeks, >20 – <= 24 weeks, and >24 weeks.

Duration of exposure (in weeks) = (Date of last study drug intake – Date of first study drug intake + 1)/7. Investigators are allowed to interrupt the study drug to allow adverse events to resolve, if necessary. Any drug interruption will not be considered when calculating the treatment duration.

Average daily dose (ug/kg/day) = Sum of (number of capsules taken per day* dose strength of the capsules per weight category * dose duration under this dose amount/weight) / (Date of last study drug intake – Date of first study drug intake + 1). Intake dose and the corresponding dose duration is collected on the dosing administration page of the eCRF.

Treatment Compliance = 100 × sum of (number of capsules actually taken) / sum of (number of capsules planned to be taken per day * dose duration under this planned dose amount). The denominator is calculated as the prescribed number of capsules (based on body weight at each visit; refer to Protocol Table 2 and 3 for the details) multiplied by the treatment period (regardless drug interrupted).

Subgroup analysis based on the Age group 1 (i.e., 0 to <10 years vs. 10 to <18 years which categorized based on the actual age) will be provided for the exposure and compliance rate. In addition, treatment compliance will be summarized quantitatively and categorically: <80%, between 80% and 120%, and >120%.

4.5 Primary Endpoint Analysis

4.5.1 Definition of Endpoint

The primary endpoint is the change from baseline in average AM and PM scratching score to Month 6 (Weeks 21 to 24) as measured by the Albireo ObsRO caregiver instrument. It is based on the worst scratching score (item 1) of the ObsRO AM and ObsRO PM scores (Protocol Appendix 2). The detailed derivation of baseline scores and monthly scores is provided in Appendix 2. Change from baseline is calculated as the monthly score minus the baseline score.

4.5.2 Estimand

The primary objective of this study is to assess the efficacy of repeated daily doses of 120 µg/kg/day odevixibat in relieving pruritus in patients with ALGS. The estimand includes all data collected through the end of the study following an ICE due to premature discontinuation of study treatment prior to Week 24 (i.e., treatment policy strategy), while excluding data following ICEs of biliary diversion surgery or liver transplant (i.e., hypothetical strategy).

On the one hand, premature treatment discontinuation could occur due to an AE, tolerability issue, lack of efficacy, or other reasons. As stated in the guidance for industry E9(R1) Statistical Principles for Clinical Trials: Addendum: Estimands and Sensitivity Analysis in Clinical Trials, the clinical and regulatory interest of hypothetical scenarios assuming the patient could tolerate treatment or did not have an AE is limited and would usually depend on a clear understanding of why and how the intercurrent event or its consequences would be expected to be different in clinical practice than in the clinical trial. Therefore, the estimand will utilize the treatment policy strategy for premature treatment discontinuation ICEs and include all data collected after this type of ICEs. This reflects the patients' outcomes in practice if they are unable to continue receiving the investigational drug. Of note, protocol section 7.1.4 specifies that for patients who prematurely discontinue study drug treatment during the treatment period, all efforts will be made to have patients continue to complete all scheduled study visits for assessments following completion of the EOT visit.

On the other hand, only a small number of patients, if any, are expected to have biliary diversion surgery or liver transplant during the placebo-controlled portion of the trial. As biliary diversion surgery may be viewed as a "rescue" treatment for pruritus, it may be reasonable for the primary analysis of pruritus outcomes to utilize a hypothetical strategy (which assumes that the biliary diversion surgery did not occur) for this ICE. A similar strategy may be reasonable for liver

transplantation, although transplantation may be indicated for reasons other than pruritus. The proposed primary analysis method will model outcomes following biliary diversion surgery or liver transplantation informed by the patients' outcomes prior to the surgery/transplant and by the trajectories of other patients in the trial.

Although not permitted by the study protocol, a significant change in anti-pruritus medication may also be considered as an ICE as it can affect the patient's scratching score and thus introduce noise into the measure of treatment effect. However, due to large number of various anti-pruritic medications with different pharmacokinetic and pharmacodynamic properties, it would be challenging from a clinical standpoint to define significant changes in anti-pruritus medication based on duration and dose collected in concomitant medication only. Therefore, change in anti-pruritus medication will not be considered as an intercurrent event in this study and all efficacy data after change in anti-pruritus medication will be included in the analyses.

4.5.2.1 Main Analysis

The primary efficacy analysis for the primary endpoint will be performed on the FAS and will be based on a MMRM excluding data after ICEs of biliary diversion surgery or liver transplant and data collected after treatment discontinuation will be included.

The change from baseline for each 4-weeks (monthly) average AM and PM scratching score will be analyzed with the SAS PROC MIXED using a restricted maximum likelihood (REML) with baseline age stratification, baseline average AM and PM scratching score, baseline direct bilirubin, treatment group, time (in months, this is a categorical variable, i.e., 1-4 weeks, 5-8 weeks, 9-12 weeks, 13-16 weeks, 17-20 weeks, and 21-24 weeks), and treatment-by-time interaction in the model. Baseline direct bilirubin is expected to be associated with the treatment effect. Unstructured covariance matrix (UN) will be used in the analysis. If this analysis fails to converge, the following structures will be tested in this order: Toeplitz (equal variances and a separate correlation for each level of separation between the time points), AR(1) (first-order autoregressive, equal variances and exponentially decreasing correlations), compound symmetry (CS) (equal variances and equal pairwise correlations across fixed time points). The first (co)variance structure yielding convergence will be used as the primary analysis. The Kenward-Roger approximation will be used to estimate denominator degrees of freedom. The primary comparison will be the difference of the least square mean (lsmean) estimate between odevixibat and placebo at Month 6 (Weeks 21 to 24). The difference of the lsmean estimate between odevixibat and placebo at other Months will be estimated and tested. Average treatment effect between odevixibat and placebo at Month 4 to 6 will be also estimated and tested.

4.5.2.2 Sensitivity Analyses

4.5.2.2.1 *Sensitivity analysis 1 – Control-based MI*

To assess the robustness of the primary analysis results based on MAR assumption, a sensitivity analysis based on MI for missing average AM and PM monthly scratching scores will be conducted.

The following procedures in MI will be followed:

1. Intermediate missing values are imputed by the Markov Chain Monte Carlo (MCMC) imputation model, so only monotone missing values exist after imputation. Only imputed values within the range of 0 to 4 will be used since the rating of the questionnaire is from 0 to 4.
2. The monotone missing values are imputed chronologically by month using SAS Proc MI regression method including baseline age stratification and baseline direct bilirubin in the model. For all patients with missing average scratching score for any reason, their missing average scratching score will be imputed from all placebo patients in a visit-by-visit (i.e., Month by Month) manner, regardless of study treatment assignment. This is done by each imputed data set from step 1. Only imputed values within the range of 0 to 4 will be used.
3. After the imputation, all data are pooled and the MMRM analysis is performed on each of the 200 imputed datasets.
4. The same MMRM model as in the primary analysis is used to compare the treatment effect at Month 6 (Weeks 21 to 24). This is done by the imputation IDs from step 1.
5. The results from step 4 will be combined using Rubin's method [14].

4.5.2.2.2 Sensitivity analysis 2 - Tipping point analysis

A tipping-point analysis will be used to explore whether missing data could have adversely impacted findings from the analysis of change from baseline in scratching score to Month 6 (Weeks 21 to 24) regardless of treatment adherence. The goal of a tipping-point analysis is to explore how the overall results may be impacted by different assumptions about the magnitude of change from baseline in scratching score to Month 6 (Weeks 21 to 24) for those with missing Month 6 (Weeks 21 to 24) data, which are sensitivity parameters (defined for each treatment group) that are systematically varied.

The tipping-point analysis will be implemented as follows:

1. 100 imputed datasets will be created using multiple imputation (MI) based on MAR, the same steps 1-2 in section 4.5.2.2.1 will be used to impute the missing average AM and PM score except that in step 2 the imputation will be done within each treatment group.
2. At each missing visit, the baseline score will be subtracted from imputed score to obtain imputed change from baseline (CHG).
3. For each patient with a missing Month 6 (Weeks 21 to 24) value, the following steps will be performed for adjusting the imputed CHG with mean zero and add the constant for the tipping steps:
    a. The average imputed CHG will be computed based on the 100 imputed datasets from step 2 per patient.
    b. The average imputed CHG will be subtracted from the imputed CHG per subject per imputed dataset and this is resulting in the imputed CHG among the imputed datasets for a subject having a mean of zero at month 6.
    c. A constant ΔO and ΔP will be added to the imputed value for the odevixibat arm and the placebo arm for each of the imputed datasets.
4. Combine imputed CHG at month 6 in step 3 and other visit in step 2 with the observed CHG to get the completed 100 datasets for analyses in step 4 and 5.

5. Results from the same MMRM model as used in the primary analysis fit to the imputed datasets will be combined using Rubin's method [14]. Treatment effect estimates, p-value and limits from the 95% confidence interval (CI) will be retained.
6. The above steps will be repeated, using different values for ΔO and ΔP.

To characterize the extent of departures from assumptions that change the interpretation of the results in terms of their statistical or clinical significance, with particular interest to identify the region in the sensitivity parameter space that leads the results to no longer be statistically significant, a map of treatment effects and p-values obtained in Step 6 above will also be presented, so as to allow assessment of the differences in posited scenarios for missing outcomes, that would lead to a change in conclusion from the study results.

4.5.2.2.3 Sensitivity analysis 3 - Worst weekly scratching score for a month

The worst weekly scratching score for baseline (14-day period) and each month (28-day period) is defined in Appendix 2; please also refer to the analysis window in Appendix 3.

The MMRM utilized for the main analysis for the sensitivity analysis of the primary endpoint will also be used for the change from baseline in scratching score to Month 6 (Weeks 21 to 24) based on worst weekly scratching score for a month. The baseline worst weekly scratching score will be used as the covariate instead of the baseline average AM and PM scratching score in the MMRM.

4.5.2.2.4 *Sensitivity analysis 4 - Worst weekly scratching score for a month with control-based MI*

The MMRM utilized for the main analysis and with control-based MI utilized for the sensitivity analysis of the primary endpoint will also be used for the change from baseline in scratching score to Month 6 (Weeks 21 to 24) based on worst weekly scratching score for a month. The baseline worst weekly scratching score will be used as the covariate instead of the baseline average AM and PM scratching score in the MMRM.

4.5.2.2.5 Sensitivity analysis 5 - Primary endpoint with different definition of baseline

The MMRM utilized for the main analysis will be used for the change from baseline in scratching score to Month 6 (Weeks 21 to 24) with the baseline score redefined as following:

The baseline scratching score for this analysis is calculated by averaging the four-baseline weekly average AM and PM scores in the 28 days preceding start of treatment for whom it is available. The baseline score will be unrounded and can only be calculated if 2 or more weekly scores can be calculated.

4.5.2.2.6 Sensitivity analysis 6 - Monthly scratching score

The MMRM utilized for the main analysis of the primary endpoint will be used for the monthly scratching score at Month 6 (week 21 to 24) to evaluate scratching at a landmark time - the end of the randomized treatment period.

4.5.2.3 Supplementary Analyses

- **MMRM for the PP analysis set**

The same MMRM utilized for the main analysis will also be used for the PP analysis set.

4.5.3 Subgroup Analyses

The following subgroup analyses will be performed by using MMRM model with baseline average AM and PM scratching score, treatment group, time and treatment-by-time interaction. The approach to selection of the covariance matrix in the model will follow the same approach as for the primary endpoint.

- Age group 1: 0 to <10 and 10 to <18 years (based on actual age)
- Age group 2: 0 to <2, 2 to <12, 12 to <18 years (based on actual age)
- Region (US, EU, and rest of the world [RoW])
- Sex
- Race (White vs. Non-White (including race as not reported))
- Ethnicity (Hispanic or Latino (including ethnicity as not reported), Not Hispanic or Latino)
- Baseline serum bile acid (≥ median and < median)
- The use of UDCA (Y, N)
- The use of anti-pruritus medication (Y, N)
- Child-Pugh classification (Class A, B, C)
- Hepatic impairment classification (no impairment, mild, moderate, or severe)
- Baseline direct bilirubin (>3 and ≤3 mg/dL and equivalent to >51.3 and ≤51.3 µmol/L)
- Genetic Testing for ALGS (JAG1 vs NOTCH2)

A forest plot of the subgroup analyses will be provided for the primary endpoint. Statistical analysis will be performed only when the sample size is ≥5 in each treatment group. If the sample size is <5 in either treatment group, only summary statistics will be provided; the p-value will not be reported.

4.6 Secondary Endpoints Analysis

4.6.1 Definition of Key Secondary Endpoint

The key secondary endpoint is change in serum bile acid levels from baseline to the average of Week 20 and Week 24. Baseline is defined as the average of the last 2 values prior to the first dose and the average of Week 20 and Week 24 is defined as the average of values at week 20 and week 24 (refer to Appendix 2 and Appendix 3 for details). Change from baseline is calculated as the serum bile acid value minus the baseline value.

4.6.2 Estimand

The key secondary objective of this study is to assess the impact of odevixibat on serum bile acid levels in patients with ALGS. The same estimand used for the primary endpoint is also used in assessing this objective. That is, all data collected through the end of the study following an ICE

due to premature discontinuation of study treatment prior to Week 24 (i.e., treatment policy strategy), will be included while excluding data following ICEs of biliary diversion surgery or liver transplant (i.e., hypothetical strategy).

4.6.2.1 Main Analysis

The analysis for the key secondary endpoint will be determined from a MMRM modelling change from baseline for serum bile acid at each scheduled visit, with baseline age stratification, baseline serum bile acid, treatment group, visits (Weeks 4, 8, 12, 16, 20, 24), and treatment-by-visit interaction in the model. The approach to selection of the covariance matrix in the model will follow the same approach for the primary endpoint. The primary comparison of treatment difference in change from baseline to the average of Week 20 and Week 24 will be estimated and tested; also, the difference at Week 4, Week 8, Week 12, Week 16, Week 20, and Week 24 will be estimated and tested.

4.6.2.2 Sensitivity Analysis

4.6.2.2.1 Sensitivity analysis 1 – Control-based MI

To assess the robustness of the main analysis results based on MAR assumption, a sensitivity analysis based on MI for missing serum bile acid result at each scheduled visit will be conducted.

The MI procedures that described for the primary endpoint in section 4.5.2.2.1 will be applied.

4.6.2.2.2 ***Sensitivity analysis 2 – Tipping point analysis***

The analysis approach that described for the primary endpoint in section 4.5.2.2.2 will be applied.

4.6.2.3 Supplementary Analysis

4.6.2.3.1 Supplementary analysis 1 – Non-parametric analysis

The serum bile acid levels are expected to fluctuate. To evaluate the robustness of the deviation from the normal and homoscedastic assumptions used in the main analysis, a rank transform ANCOVA model with baseline age stratification, baseline serum bile acid, treatment group will also be used to evaluate the treatment effect. The ranked ANCOVA method permits adjustment for baseline values which increases the efficiency of the statistical test. Change from baseline to the average of Week 20 and Week 24 will be ranked ignoring treatment group. The value of mid-ranks will be used in the case of ties. The baseline serum bile acid data will also be ranked in the same manner. This analysis will be performed with missing data at each scheduled visit imputed by MI based on MAR (the step 1 in the tipping point analysis) before ranking. The multiple estimates of treatment difference from the ranked ANCOVA will then be combined using Rubin's rules. The same ranked ANCOVA will also be performed at each visit (Weeks 4, 8, 12, 16, 20, 24).

4.6.2.3.2 *Supplementary analysis 2 – Based on the PP analysis set*

The MMRM utilized for the main analysis in section 4.6.2.1 will also be used for the PP analysis set.

4.6.3 Subgroup Analyses

The following subgroup analyses will be performed by using MMRM model with baseline serum bile acid, treatment group, visit and treatment-by-visit interaction. The approach to selection of the covariance matrix in the model will follow the same approach as for the primary endpoint.

- Age group 1: 0 to <10 and 10 to <18 years (Based on the actual age)
- Age group 2: 0 to <2, 2 to <12, 12 to <18 years (Based on the actual age)
- Region (US, EU, and RoW)
- Sex
- Race (White vs. Non-White (including race as not reported))
- Ethnicity (Hispanic or Latino (including ethnicity as not reported), Not Hispanic or Latino)
- Baseline serum bile acid (≥ median and < median)
- The use of UDCA (Y, N)
- The use of anti-pruritus medication (Y, N)
- Child-Pugh classification (Class A, B, C)
- Hepatic impairment classification (no impairment, mild, moderate, or severe)
- Baseline direct bilirubin (>3 and ≤3 mg/dL and equivalent to >51.3 and ≤51.3 µmol/L)
- Genetic Testing for ALGS (JAG1 vs NOTCH2)

A forest plot of the subgroup analyses will be provided for the key secondary endpoint. Statistical analysis will be performed only when the sample size is ≥5 in each treatment group. If the sample size is <5 in either treatment group, only summary statistics will be provided; the p-value will not be reported.

4.6.4 Other Secondary Endpoints

For all secondary endpoints, all data collected through the end of the study following an ICE due to premature discontinuation of study treatment prior to Week 24 (i.e., treatment policy strategy), will be included while excluding data following ICEs of biliary diversion surgery or liver transplant (i.e., hypothetical strategy).

The change in secondary endpoints such as serum bile acids, ALT, AST, gamma-glutamyl transferase, direct bilirubin, total bilirubin, total cholesterol, PRO and ObsRO sleep parameters will be summarized by visit using descriptive statistics.

In addition to evaluate vital signs (height and weight) in the safety analysis, growth data (height/length for age, weight for age, and BMI for age) will be considered as an exploratory efficacy endpoint. Growth data will be calculated using the software or methods from the Centers for Disease Control and Prevention (CDC) website [12] for patients with age ≥2 years old and from the World Health Organization (WHO) website [13] for patients with age <2 years old.

Change from baseline in growth (i.e., height [cm], weight [kg], BMI [kg/m^2] and corresponding linear growth compared to a standard growth curve [Z-score, SD from P50 standard growth curve], calculated by using the software or methods from the CDC website for patients with age >= 2 years old and from the WHO website for patients with age < 2 years old) will be summarized by visit using descriptive statistics.

4.6.4.1 Continuous Endpoints

Change in ALT, AST, gamma-glutamyl transferase, direct, total bilirubin and total cholesterol, change in itching score to Month 6 (Weeks 21 to 24) as measured by the Albireo PRO instrument, change in sleep parameters, and change in growth will also be analyzed using the MMRM model. The summary of change from baseline based on MMRM model will also be displayed using graphical presentations except for itching score as measured by the Albireo PRO instrument. MMRM model will include the baseline value of the response variable, baseline age stratification, baseline direct bilirubin (for change in itching score), treatment group, visit (in months for itching score and sleep parameters, and in weeks for other assessments) and treatment-by-visit interaction. The approach to selection of the covariance matrix in the model will follow the same approach as for the primary endpoint.

The change from baseline through Week 24 in the Albireo ObsRO scratching severity scores and PRO itching scores for the morning assessment, evening assessment, and average morning and evening assessment will also be evaluated by weekly score. These endpoints will be assessed combining age groups and by age group (based on the actual age), 0 to <8, 8 to <12, and 12 to <18. A line graph of scratching daily severity scores (average AM and PM) from the Albireo ObsRO and sBA over time for each patient will be provided. In addition, line plot for mean of change from baseline on Albireo ObsRO average AM and PM, AM, and PM scratching weekly score will be provided, respectively.

Patients (≥5 years) and caregivers will be asked to complete a quality-of-life questionnaire (PedsQL) at Randomization Day 1, Week 12, and Week 24. Details of the questions included on the questionnaire are provided in Protocol Appendix 4. Comparisons of the change from baseline at Week 12 and Week 24 in the PedsQL total score (calculated as average score of all answered items) and domain scores between the treatment groups will be conducted using ANCOVA. The model will include terms for baseline, age category based on the age groups defined for the PedsQL (5 to <8, 8 to <13, 13 to <18) and treatment. The analysis will be conducted based on the total scores reported by child (≥5 years of age) and by parent (including patients for all age) separately. If the sample size based on child reported data (≥5 years of age) is less than 10, then ANCOVA will not be conducted. The PedsQL total score and domain scores of the family impact module will be analyzed similarly.

A Pearson correlation coefficient will be provided to evaluate the relationship between the primary endpoint (change from baseline in scratching score to Month 6 (Weeks 21 to 24)) and the key secondary endpoint (change in serum bile acid levels from baseline to the average of Week 20 and Week 24).

4.6.4.2 Categorical Endpoints

A clinically meaningful decrease in scratching score will be defined in the psychometric analysis report based on psychometric analysis results before database lock. Please refer to the

psychometric analysis plan for details. The results of the blinded psychometric analysis across all anchors and timepoints supported a threshold from 1.0 to 1.5 points. The upper bound of 1.5 points reduction will be used for the primary analysis while the lower bound of 1.0-point reduction will be used for the sensitivity analysis. The proportion of patients achieving a clinically meaningful decrease in scratching score (pruritus responders) at Week 24 (or Week 12) will be analyzed using a Cochran-Mantel-Haenszel (CMH) test. Patients with missing data at Week 24 (or Week 12) will be classified as non-responders. CMH stratified by baseline age stratification will be performed to compare the proportion of the responders in pruritus at Week 24 (or Week 12). In the CMH test, data in a stratum will not be used in the calculation of the p-value if a row sum or column sum is 0 in the contingency table. If a row sum or column sum is 0 or a stratum has <4 patients, the CMH test will not be stratified.

The clinician xanthoma scale uses a 5-point scale, in which 0 represents no evidence of xanthomatosis, 1 represents fewer than 20 scattered individual lesions, 2 represents more than 20 lesions that do not interfere with or limit activities, 3 represents large numbers of lesions that by their large numbers or size cause distortion of the face or extremities, and 4 represents xanthomas that interfere with function (such as hand use or ability to walk) because of excess size or number. The change from baseline to Week 12, Week 24 and last assessment during treatment in xanthomatosis will be summarized by a shift table. Shift tables for the number of patients with changes from baseline to the best and to the worst post-baseline value will be provided.

Patients (≥8 years of age), caregivers, and clinicians will complete GIC (PGIC, CaGIC, and CGIC) and the GIS measures (PGIS, CaGIS, CGIS) (Protocol Appendix 3) at randomization (Study Day 1; PGIS only), Week 4, 12, and 24. The GIC items that assess change in itch (patient version), scratching (caregiver and clinician version), and sleep (all version) will be analyzed using the proportional odds model with treatment and baseline GIS score as the covariates and will consider 3 categories (better, no change, worse). The better category will include answers of “Very Much Better”, “Much Better”, and “A Little Better”; no change will include answers of “No Change”; worse will include answers of “A Little Worse”, “Much Worse”, and “Very Much Worse”. GIC and GIS data at baseline (GIS only), Weeks 4, 12, and 24 will be summarized using frequency tables and bar chart figures.

4.6.4.3 Biomarkers

Blood samples for autotaxin and p-C4 will be drawn at randomization (Study Day 1), Week 12, and Week 24, only for children with body weight >10 kg. For biochemical markers and measures of bile acid synthesis, the change from baseline to Week 24 will be analyzed descriptively in addition to the actual visit values.

FGF-19 is collected when a patient meets criteria for close observation due to suspected drug-induced liver injury (DILI); the results will be provided in a listing.

4.6.4.4 Exit Interview

During the exit interview, patients (≥8 years of age) and caregivers will be asked to complete the exit questionnaire (Protocol Appendix 8). The answers for the first two questions on the questionnaire will be summarized using a frequency table for each combination of the answer from these two questions by the responses from patients and caregivers. The details of responses to the exit interview will be provided in a listing.

4.7 Multiplicity Adjustment for Primary and Key Secondary Endpoints

Statistical testing for the primary analysis of the primary endpoint will be performed with a 1-sided Type I error rate of 0.025. The key secondary efficacy endpoint will be assessed for statistical significance if and only if the success criterion for the primary endpoint is met. This hierarchical testing strategy prevents alpha inflation from the assessment of multiple endpoints. Other secondary efficacy endpoints will provide supportive efficacy information regarding the differences between treatment with odevixibat and placebo. No alpha adjustments will be performed for multiple comparisons when testing other secondary efficacy endpoints.

4.8 Safety Analyses

Safety criteria are as follows:

- Primary safety analysis is the incidence of TEAEs and TEAEs categorized by causality, severity, and seriousness assessments made by the investigator by comparing active arms vs. placebo. This includes liver-related mortality and liver decompensation events (protocol Section 10.2.2.3) and all-cause mortality.
- Safety will also be evaluated by the following assessments:
  - Physical examinations
  - Concomitant medications
  - Vital signs
  - Laboratory test results (including clinical chemistry, hematology, urinalysis, AFP, vitamins A and E, 25-hydroxy vitamin D, and INR)
  - Abdominal ultrasound
  - Discontinuations due to AEs

All safety analyses will be performed on the SAF unless otherwise specified, using the following common rules:

- The analysis of the safety variables will be analyzed using descriptive statistics and summaries by treatment group for each cohort.
- No inferential statistical analysis will be planned.
- For each safety variable, the last value collected prior to first dose of study drug will be used as the baseline for all analyses.

4.8.1 Adverse Events

General common rules for summarizing adverse events

All AEs will be classified by SOC and PT using the MedDRA version 25.0.

AEs will be classified as TEAEs, defined as follows: An AE (classified by preferred term) that occurs during the Treatment Period will be considered a TEAE if it has a start date on or after the first dose date of study drug, or it has a start date before the date of the first dose date of study drug but worsened in severity on or after the date of the first dose date of study drug.

AEs with missing start dates, but with stop dates either overlapping with the treatment period or missing, will be considered TEAEs. A TEAE with missing drug-relationship will be considered as related. A TEAE with missing severity will be considered as CTCAE grade 3 (severe). Details for imputing missing or partial start dates of adverse events are described in Appendix 5. AEs with the same AE reference number that is collected on AE page of eCRF are counted as one AE only.

Adverse events of interest (AEIs)

Adverse events of interest (AEIs) will be selected for analyses as indicated in Table 6 - Selections for AEIs .

Table 6 – Selections for AEIs

AEIs	Selection
Liver Decompensation	a. The following TEAEs will be included: PT = International normalised ratio increased and checkbox that event is refractory to vitamin supplementation is checked and at least 1 INR value (within 2 weeks [+/-] of the AE started date) is >1.5 PT = Vitamin K increased and checkbox that event is refractory to vitamin supplementation is checked and at least 1 INR value (within 2 weeks [+/-] of the AE started date and checkbox) is >1.5 PT = Hepatic failure, Acute on chronic liver failure, or Subacute hepatic failure b. The following TEAEs occurring as separate events or in the presence of (i.e. ongoing medical history or AE) PT of portal hypertension and/or PT of hepatic cirrhosis were included: PT = Ascites PT = Hepatorenal syndrome PT = Portopulmonary hypertension PT = Oesophageal varices haemorrhage PT = Gastric varices haemorrhage PT = Intestinal varices haemorrhage PT = Hepatic encephalopathy
Liver Related	The following Standardised MedDRA Queries (SMQs) will be included: • Drug related hepatic disorders – comprehensive search SMQ (narrow and broad) • Biliary tract disorders SMQ • Gallbladder related disorders SMQ • Gallstone related disorders SMQ
FSV Deficiency refractory to clinically recommended vitamin supplementation based on CRF	Both questions of “Is this an event of fat-soluble vitamin deficiency?” and “Is this refractory to clinically recommended vitamin supplementation?” on the AE page are checked as “Yes”.
Possible Sequelae of FSV Deficiency	Specified TEAEs with PTs are listed in Appendix 10

AEIs	Selection
Clinically Significant Diarrhea	Clinically significant diarrhea is defined as diarrhea that persists for 3 or more days without any other etiology (e.g., infection) as assessed by the investigator, diarrhea reported by the investigator as severe in intensity or reported as an SAE or diarrhea with concurrent dehydration requiring treatment with oral or intravenous rehydration and/or other treatment intervention. Therefore, at least one of the following conditions met will be considered as clinically significant diarrhea: • PT = Diarrhoea or Diarrhoea haemorrhagic, and last for ≥3 days and the CRF question of "Does this event have a clear etiology (e.g., Infection)?" on the AE page is checked as "No". • PT = Diarrhoea or Diarrhoea haemorrhagic and CTCAE ≥G3 • PT = Diarrhoea or Diarrhoea haemorrhagic and SAE = "Yes" • PT = Diarrhoea or Diarrhoea haemorrhagic and the CRF question of "Is this an event of diarrhea with concurrent dehydration requiring treatment with oral or intravenous rehydration and/or other treatment intervention?" on the AE page is checked as "Yes"

Analysis of all adverse events

An overall summary of the incidence of TEAEs (number of patients with any events and number of events, if applicable) will include the following:

- Overview of the incidence of TEAEs
- Drug-related TEAEs (AE will be defined as drug-related if causality is either probable, possibly, or definitely related)
- TEAEs by CTCAE grades 1, 2, 3, 4, 5
- TEAEs leading to study drug discontinuation
- TEAEs leading to study drug interruption
- TEAEs leading to study drug dose reduction
- Serious TEAEs
- Drug-related serious TEAEs
- TEAEs leading to death
- Any hepatic TEAE
  - Any TEAE of liver decompensation (as defined in Table 6 - Selections for AEIs, for TEAEs listed in b occurring as separate events)
  - Any TEAE of liver decompensation (as defined in Table 6 - Selections for AEIs, for TEAEs listed in b in the presence of (i.e. ongoing medical history or AE) PT of portal hypertension and/or PT of hepatic cirrhosis)
  - Any liver-related TEAE (as defined in Table 6 - Selections for AEIs)
- Any Fat-Soluble Vitamin (FSV) Deficiency refractory to clinically recommended vitamin supplementation TEAEs (as defined in Table 6 - Selections for AEIs)
- Possible Sequelae of FSV Deficiency TEAEs (as defined in Table 6 - Selections for AEIs)

- Any Clinically Significant Diarrhea TEAEs (as defined in Table 6 - Selections for AEIs)
- Overview of the incidence of TEAEs by age group 1 (0 to <10, 10 to <18, based on the actual age)
- Overview of the incidence of TEAEs by age group 2 (0 to <2, 2 to <12, 12 to <18, based on the actual age)

TEAEs (number of patients with any events and number of events, if applicable) by SOC (as applicable) and PT will be tabulated for the following:

- TEAEs by SOC and PT
- All TEAEs by PT by descending incidence in odevixibat column
- TEAEs leading to study drug discontinuation by SOC and PT
- TEAEs leading to study drug interruption by SOC and PT
- TEAEs leading to study drug dose reduction by SOC and PT
- Serious TEAE by SOC and PT
- Drug-related TEAEs by SOC and PT
- Drug-related serious TEAE by SOC and PT
- TEAEs leading to death by SOC and PT
- Common TEAEs (≥5% odevixibat group) by PT
- Severe or worse of TEAEs (CTCAE grade ≥ 3) by SOC and PT
- Hepatic TEAEs
  - Any TEAE of liver decompensation (as defined in Table 6 - Selections for AEIs, for TEAEs listed in b occurring as separate events) by SOC and PT
  - Any TEAE of liver decompensation (as defined in Table 6 - Selections for AEIs, for TEAEs listed in b in the presence of (i.e. ongoing medical history or AE) PT of portal hypertension and/or PT of hepatic cirrhosis) by SOC and PT
  - Liver-related TEAEs by SOC and PT
- FSV Deficiency refractory to clinically recommended vitamin supplementation TEAEs by SOC and PT (as defined in Table 6 - Selections for AEIs)
- Possible Sequelae of FSV Deficiency TEAEs by SOC and PT (as defined in Table 6 - Selections for AEIs)
- Clinically Significant Diarrhea TEAEs by SOC and PT (as defined in Table 6 - Selections for AEIs)

Summary tables for number of patients with any TEAEs by SOC and PT by severity (grade 1 to grade 5) and by causality (related [possibly, probably, and definitely] vs unrelated [unlikely and unrelated]), will also be provided. AEs with the worst severity will be used in the by-severity summaries. Similarly, AEs with the worst causality (most related to treatment) will be used in the by-causality summaries. If severity or causality is missing, data will be imputed as grade 3 or related, respectively.

When a patient has the same adverse event, based on preferred terminology, reported multiple times during the study, the patient will only be counted once at the preferred terminology level in adverse event frequency tables. Where a patient has multiple adverse events within the same SOC during the study, the patient will only be counted once at the SOC level in adverse event frequency tables. In the AE summaries, AEs will be sorted by alphabetically for SOC and PT.

In addition to the frequency summary of TEAE tables, the exposure adjusted incidence rate (EAIR) (per 100 patient years at risk) with the 95% confidence interval, based on Poisson distribution, will also be provided for each of the following AE tables:

- TEAEs by SOC and PT
- Liver-related TEAEs by SOC and PT
- Liver Decompensation TEAEs (as defined in Table 6 - Selections for AEIs, for TEAEs listed in b in the presence of (i.e. ongoing medical history or AE) PT of portal hypertension and/or PT of hepatic cirrhosis) by SOC and PT
- FSV Deficiency refractory to clinically recommended vitamin supplementation TEAEs by SOC and PT
- Possible Sequelae of FSV Deficiency TEAEs by SOC and PT
- Clinically Significant Diarrhea TEAEs by SOC and PT

The EAIR (per 100 patient years at risk) will be calculated as follows:

Number of subjects with event/ Total cumulative time at risk (years) x 100

where total cumulative time at risk is the sum of individual time at risk. The time at risk for a subject with the event of interest will be calculated (in years) as:

(first event date – date of first dose + 1) / 365.25

The time at risk for a subject without the event of interest is calculated as:

(last participation date – date of first dose + 1) / 365.25

Kaplan-Meier curves may be used for time to event data (time to first liver decompensation TEAEs [as defined in Table 6 - Selections for AEIs, for TEAEs listed in b in the presence of (i.e. ongoing medical history or AE) PT of portal hypertension and/or PT of hepatic cirrhosis] onset (days), time to first liver-related TEAEs onset (days), time to first FSV deficiency refractory to clinically recommended vitamin supplementation TEAEs onset (days), time to first possible sequelae of FSV deficiency TEAEs onset (days), time to first clinically significant diarrhea event onset (days), and time to all-cause mortality (weeks)). Patient will be censored at the last contact date if no event occurs.

A diarrhea event is defined as a TEAE with PT as Diarrhoea or Diarrhoea haemorrhagic. Descriptive summary statistics will be used to summarize the duration of each diarrhea episodes (days), the number of episodes per patient and time to onset of diarrhea (days). In addition, the summaries will be provided for clinically significant diarrhea TEAEs.

In order to assess who developed onset of FSV Deficiency refractory to clinically recommended vitamin supplementation, a listing merging with concomitant medications and selected lab data will be provided.

Analysis of special situations

Certain safety events are defined by regulatory authorities as Special Situations and require reporting. These Special Situations include:

- Overdose of study medication
- Suspected abuse/misuse of study medication
- Medication error involving the study medication
- Drug-drug interaction
- Pregnancy

Special Situations should be reported on the Special Situations page of eCRF, whether or not they result in an AE/SAE. Number (%) of patients experiencing these special situations will be provided.

4.8.2 Additional Safety Assessments

4.8.2.1 Laboratory Variables

The following laboratory variables will be analyzed.

- Hematology:
  - Red blood cells and platelets: hemoglobin, hematocrit, red blood cell count, and platelet count
  - White blood cells: white blood cell count, neutrophils, lymphocytes, monocytes, basophils, and eosinophils
- Clinical chemistry:
  - Metabolism: total cholesterol, triglycerides, creatine kinase, albumin
  - Electrolytes: sodium, potassium, chloride, calcium
  - Renal function: creatinine, blood urea nitrogen, estimated glomerular filtration rate
  - Liver function: ALT, AST, Alkaline phosphatase (and isoenzymes as applicable), direct bilirubin, total bilirubin, gamma-glutamyl transferase
  - Pregnancy test: Serum β-human chorionic gonadotropin (all female patients with positive urine pregnancy test)
- Urinalysis:
  - Urinalysis for quantitative analysis: pH, proteins, glucose, blood, ketones, leukocytes, and nitrites
- Other Labs: Vitamin A and E, 25-hydroxy vitamin D, INR, Alfa-fetoprotein (AFP), and Prothrombin Time

Quantitative analyses

For all laboratory variables above, descriptive statistics for clinical laboratory values (in SI units for all tests, and in conventional units for selected tests (Appendix 8) and absolute changes from baseline at each post-baseline visit (Appendix 3) will be presented. The change to the last visit will also be summarized. Central laboratory data will be used for the summary. If central laboratory data are not available due to COVID-19 or other reasons, local laboratory data will be used if available.

Figure for ALT, AST, total bilirubin, ALP, and gamma-glutamyl transferase will be presented as individual patient trellis plot over time.

Analyses according to reference normal range

Shift tables from baseline to the highest and lowest post-baseline value for quantitative variables will be presented for the clinical laboratory parameters. The laboratory results will be classified as low, normal, and high based on the normal range that provided by central/local laboratories.

For ALT, AST and total bilirubin, a shift table based on the following classifications (per DILI guidance and CTCAE) will be presented:

- ALT and AST: ≤ 3 ULN, > 3 and ≤ 5 ULN, > 5 and ≤ 10 ULN, > 10 ULN
- Total Bilirubin: ≤ 2 ULN and > 2 ULN

Liver monitoring

A listing of patients meeting the criteria for liver monitoring (Table 7) will be provided.

Table 7 – Liver monitoring

Laboratory value(s) or Symptoms	Criterion
AST or ALT in combination with total bilirubin (confirmed with repeat testing within 48 to 72 hours)	CCI
Total bilirubin (confirmed with repeat testing within 48 to 72 hours)	
INR	

Laboratory value(s) or Symptoms	Criterion
Symptoms	CCI [redacted]
Symptoms	

4.8.2.2 Vital Signs

Descriptive statistics for vital signs [blood pressure (systolic and diastolic), pulse, respiratory rate, temperature, height (or length depending on age), body weight (using a certified weight scale), and BMI] and their changes from baseline at each post-baseline visit will be presented. The change to last visit will also be summarized. Shift tables for the number of patients with changes from baseline to highest and lowest post-baseline value based on the normal ranges as outline in Appendix 7 will be provided.

If height is collected as length, it will be converted as following:

Height (cm) = length (cm) – 0.7 cm.

4.9 Other Analyses

4.9.1 Pharmacokinetics (PK) Analyses

PK analysis will be specified in a separate PK SAP.

4.9.2 Blinded Analysis

4.9.2.1 Pruritus Responder Definitions

Blinded analyses of the Albireo ObsRO (or PRO) eDiary data will be performed when approximately 85% or more of the planned sample size has completed the Week 24 (or Week 12) visit. The blinded analysis will be used to estimate a threshold of clinically meaningful change (i.e., responder definition) in the Albireo ObsRO scratching score (or PRO itching score). The analysis will be performed by a group from IQVIA, independent from both the study team and Albireo. Blinded data will be used for this analysis; data will be collapsed across treatment groups.

The analyses will include distribution- and anchor-based analyses to examine a responder definition for the Albireo ObsRO scratching score (or PRO itching score). Distribution-based analyses will include calculation of the SD and standard error of measurement for the baseline

Albireo ObsRO scratching score (or PRO itching score). Anchor-based analyses will involve examining the degree of change on the Albireo ObsRO scratching score (or PRO itching score) from baseline to Week 24 (or Week12) for patients who experience change in pruritus according to PGIC and symptom items. This will be described in the PAP.

5 Supporting Documentation

5.1 Appendix 1 Changes to Protocol-Planned Analyses

This section summarizes major statistical changes in the protocol amendment(s).

Table 8 - Major statistical changes in protocol amendment(s)

Amendment Number	Approval Date	Changes	Rationale
V2.0	28Aug2020	Definition of the PP analysis set is changed for patients without **major** protocol violation to patient without **important** protocol violation. Additional criteria are added.	Correct the definition of the PP analysis set.
		The covariate of baseline conjugated bilirubin in MMRM is changed to direct bilirubin.	Align with the routine laboratory parameters collected during the study
		The contrast statement for month 6 (weeks 21 to 24) to assess treatment effect is changed to the difference of the least square mean estimate.	The difference of the least square mean estimate will produce the same result as the contrast statement for the primary endpoint at Month 6 and provide data for the preceding visits.
		Add growth as an exploratory efficacy endpoint	Patients with ALGS have impaired growth. To evaluate the effect of odevixibat on growth.

5.2 Appendix 2 Derived Variables

The table below provides the list of derived variables for demographic and baseline characteristics, various duration derivations, drug compliance, baseline derivations and other important protocol derivations applicable for this study.

Table 9 – Derivation rules for the derived variable

Variables	Formula
Demographic and Baseline characteristics	
Age (in years)	For patients not from site CCI : Age is calculated based on date of birth and date of randomization. It will be derived by SAS YRDIF function: yrdif (Date of birth, Date of Randomization, "ACT/ACT") For patients from site CCI : Only year of birth was collected, and January 1 was imputed in the eDC system. For analysis purpose, age is calculated based on collected age months and age years on the CRF.
Derivation of Duration	
Study day at any visit	Date of interest – Date of first dose of study drug. One day is added if this difference is ≥0
Extent of Exposure	
Duration of treatment exposure (days)	Date of last study drug intake – Date of fist study drug intake+1
Average Daily dose (ug/kg/day)	Sum of (number of capsules taken per day * dose strength of the capsules per weight category * dose duration under this dose amount/weight) / (Date of last study drug intake – Date of first study drug intake + 1)
Overall Study Drug Compliance	
Compliance	100 × sum of (number of capsules actually taken) / sum of (number of capsules planned to be taken per day * dose duration under this planned dose amount)
Last participant date	Last contact date or date of death which occurred first
Derivation for Efficacy Parameters	
Baseline (general)	The baseline value of a parameter is defined as the last assessment of that parameter prior to the first intake of study treatment, except as indicated below
Baseline for serum bile acid	The baseline will be calculated as the average of the last 2 values prior to the first dose (central laboratory results will be used over local laboratory results). If only one non-missing value is available, it will be used as baseline.
Average serum bile acid of Week 20 and Week 24	The average of values that identified from analysis window at Week 20 and Week 24.
Daily score for ObsRO scratching (or PRO itching) severity	The daily score is calculated by averaging the PM score and the next day AM score. The daily score is calculated if at least the PM score or the next day AM score is collected.

Variables	Formula
Weekly AM (or PM) score for ObsRO scratching (or PRO itching) severity	The weekly AM (or PM) score is calculated by averaging the AM (or PM) scores in a week. Each week, at least 4 of 7 AM (or PM) assessments need to be collected to calculate the weekly AM (or PM) score (i.e., 50% rule). If these minimum assessments are not available, the weekly AM (or PM) score is considered missing
Weekly average score for ObsRO scratching (or PRO itching) severity	The weekly average AM and PM score is calculated by averaging of (the average of AM scores, the average of PM scores) in a week. Each week, at least 7 of 14 AM or PM assessments need to be collected to calculate the weekly average AM and PM score (i.e., 50% rule) regardless how many AM assessments or how many PM assessments are collected respectively. If these minimum assessments are not available, the weekly average AM and PM score is considered missing.
Baseline score for ObsRO scratching (or PRO itching) severity (general)	The baseline AM (or PM, or average AM and PM) score is calculated by averaging the two-baseline weekly AM (or PM, or average AM and PM) scores in the 14 days preceding start of treatment. Baseline score will be unrounded. Baseline values can only be calculated if both weeks can be calculated.
Baseline score for ObsRO scratching (or PRO itching) severity (sensitivity analysis)	The baseline AM (or PM, or average AM and PM) score is calculated by averaging the four-baseline weekly AM (or PM, or average AM and PM) scores in the 28 days preceding start of treatment. Baseline score will be unrounded. Baseline values can only be calculated if 2 or more weeks score can be calculated.
Monthly average score for ObsRO scratching or (PRO itching) severity	The post-baseline monthly (28 days) AM (or PM, or average AM and PM) score is calculated by averaging 4 weekly scores within the 4-week interval. The monthly AM (or PM, or average AM and PM) score can only be calculated if at least 3 of 4 weekly scores within the 4-week interval can be calculated.
Baseline for worst weekly score for ObsRO scratching (or PRO itching) severity	The baseline for worst weekly score is calculated by taking the worst daily score as the maximum of the PM score and the next day AM score in the 14 days preceding start of treatment, then calculating the worst weekly score by taking the average of the worst daily scores in a week (7 days), and finally calculating the worst weekly score for baseline as the maximum of the worst weekly scores within the 2-week interval. At least one of the PM score or the next day AM score need to be collected to calculate the worst daily score. At least 4 of 7 worst daily scores within the week are needed to calculate the worst weekly score. The worst weekly score for baseline is calculated if at least one worst weekly score within the 2-week interval can be calculated.
Worst weekly score for a month for ObsRO scratching (or PRO itching) severity	The worst weekly score for a month (28 days) is calculated by taking the worst daily score as the maximum of the PM score and the next day AM score, then calculating the worst weekly score by taking the average of the worst daily scores in a week (7 days), and finally calculating the worst weekly score for a month as the maximum of the worst weekly scores within the 4-week interval. At least one of the PM score or the next day AM score need to be collected to calculate the worst daily score. At least 4 of 7 worst daily scores within the week are needed to calculate the worst weekly score. The worst weekly score for a given month is calculated if at least one worst weekly score within the 4-week interval can be calculated. The worst weekly score for a month is calculated for each month of post-baseline

Variables	Formula
Change from baseline for ObsRO scratching or (PRO itching) severity score	The daily, weekly, monthly AM (or PM, or average AM and PM) score change from baseline is calculated by subtracting the baseline AM (or PM, or average AM and PM) score from the daily, weekly, monthly AM (or PM, or average AM and PM) score
Change from baseline for worst weekly score for a month for ObsRo scratching (or PRO itching) severity	Change from baseline in the worst weekly score for a month is calculated by subtracting the worst baseline score from the worst weekly score for a month
Average Month 4 to 6 and change from baseline	Average of all available monthly scores through Month 4 to 6. Change from baseline through Month 4 to 6 is calculated by subtracting the baseline score from the average of monthly scores through Month 4 to 6.
Baseline and monthly score for sleep parameters (such as difficulty of falling asleep and staying asleep, tiredness, and the number of awakening)	For patient- and observer- reported outcome scores of sleep parameters, the same approach above will be used as for scratching (or itching) severity score by AM and PM respectively since there is just 1 rating per day.
PRO and ObsRO Assessment Scores for psychometric analyses	Please refer to psychometric analysis plan

Derivation for Safety Parameters	
AE duration (days)	AE end date – AE start date +1
TEAE	An AE (classified by preferred term) that occurs during the Treatment Period will be considered a TEAE if it has a start date on or after the first dose date of study drug, or it has a start date before the date of the first dose date of study drug, but worsened in severity on or after the date of the first dose date of study drug

5.3 Appendix 3 Data Handling Conventions

Analysis windows for time points

The following analysis windows will decide how the scheduled and/or unscheduled visits will be used in the by-visit analyses of efficacy and safety variables. A measurement (scheduled or unscheduled) will be used if it is available and measurement date is within the analysis window.

After applying these time windows, if multiple assessments are associated to the same time point, the closest from the targeted study day will be used. In case of ties between observations located on different sides of the target day, the earlier assessment will be used. For laboratory parameters, central laboratory results will be used over local laboratory results if both are in a same analysis window.

Table 10.1 - Analysis visit windows (general)

Timing of assessment	Visit Name to display for Analysis	Targeted study day	Analysis window in study days / date of assessment
Screening	Baseline		<= -1
Randomization day 1	Baseline		1 (pre-dose)
Week 4	Week 4	28	Post-baseline to 42
Week 8	Week 8	56	43 to 70
Week 12	Week 12	84	71 to 98
Week 16	Week 16	112	99 to 126
Week 20	Week 20	140	127 to 154
Week 24	Week 24	168	155 to EOS[a] / 155 to EOS-1[b]
4 Weeks Post Last Dose of Study Drug	Follow-up		EOS[b]

EOS=date of EOS assessment
[a] For patients who rollover to Study A4250-015.
[b] For patients who not rollover to Study A4250-015.

Table 10.2 - Analysis visit windows (GIC/GIS, PedsQL, xanthoma and selected lab tests)

Timing of assessment	Visit Name to display for Analysis	Targeted study day	Analysis window in study days / date of assessment
Screening	Baseline		<= -1
Randomization day 1	Baseline		1 (pre-dose)
Week 4	Week 4	28	Post-baseline to 56
Week 12	Week 12	84	57 to 126
Week 24	Week 24	168	127 to EOS[a] / 127 to EOS-1[b]
4 Weeks Post Last Dose of Study Drug	Follow-up		EOS[b]

Note Selected lab tests include hematology, urinalysis, autotaxin, p-C4, Vitamins A & E and 25-hydroxy vitamin D.
EOS=date of EOS assessment
[a] For patients who rollover to Study A4250-015.
[b] For patients who not rollover to Study A4250-015.

Table 10.3 - Analysis visit windows (abdominal ultrasound and elastography, AFP)

Timing of assessment	Visit Name to display for Analysis	Targeted study day	Analysis window in study days / date of assessment
Screening	Baseline		<= -1
Randomization day 1	Baseline		1 (pre-dose)
Week 24	Week 24	168	71 to EOS[a] / 71 to EOS-1[b]
4 Weeks Post Last Dose of Study Drug	Follow-up		EOS[b]

EOS=date of EOS assessment
[a] For patients who rollover to Study A4250-015.
[b] For patients who not rollover to Study A4250-015.

Derivations for scratching score (or itching score), sleep parameters measured by the Albireo ObsRO (or PRO) instrument will be derived based on the following analysis window. For analysis purposes, diary entries will be assigned to a study day based on the recorded date regardless of recorded time.

Table 10.4 - Analysis visit windows for scratching score (or itching score), sleep parameters

Analysis Window	Treatment Period	Start Study Day	End Study Day
Baseline	Week 0 (randomization)	Day -14, PM assessment Day -28, PM assessment [a]	Day 1, AM assessment
Week 1	Week 1	Day 1, PM assessment	Day 8, AM assessment
Week 2	Week 2	Day 8, PM assessment	Day 15, AM assessment
Week X	Week X	Day 7*(x-1)+1, PM assessment	Day 7*x+1, AM assessment

[a] Day -28 is for sensitivity analysis.

5.4 Appendix 4 Definition of Region Variable

Table 11 – Country code and region

Country	Country code	Region variable
Belgium	BEL	EU
Canada	CAN	RoW
France	FRA	EU
Germany	DEU	EU
Israel	ISR	RoW
Italy	ITA	EU
Malaysia	MYS	RoW
Netherlands	NLD	EU
New Zealand	NZL	RoW
Poland	POL	EU
Turkey	TUR	RoW
United Kingdom	GBR	RoW
United States	USA	US

5.5 Appendix 5 Handling of Missing or Incomplete Dates

Global statement:
If the imputed date is prior to the date of birth, then impute the missing date as date of birth.

Imputation rules for missing or partial AE start date are defined below:

If only Day of AE start date is missing:
If the AE start year and month are the same as that for the first dose date, then:

- If the full (or partial) AE end date is NOT before the first dose date or AE end date is missing, then impute the AE start day as the day of first dose date;
- If AE end date is prior to first dose date, then impute the AE start day as 1.

If Day and Month of AE start date are missing:
If AE start year = first dose year, then:

- If the full (or partial) AE end date is NOT before the first dose date or AE end date is missing, then impute the AE start month and day as the month and day of first dose date;
- If AE end date is prior to first dose date, then impute the AE start month as January and the day as 1.

If Year of AE start date is missing:
If the year of AE start is missing or AE start date is completely missing, then query the site with no imputation. Also compare the full (or partial) AE end date to the first dose date. If the AE end date is before the first dose date, then the AE should be considered as a pre-treatment AE. Otherwise, the AE will be considered as TEAE.

Imputation rules for missing or partial non-study medication start/stop dates are defined below:

Missing or partial non-study medication start date:

- If only day is missing, use the first day of the month.
- If day and month are both missing, use the first day of the year.
- If day, month and year are all missing, use the date of the day before the first dose date.

Missing or partial non-study medication stop date:

- If only day is missing, use the last day of the month.
- If day and month are both missing, use the last day of the year.
- If day, month and year are all missing, assign ‘continuing’ status to stop date

5.6 Appendix 6 Sample SAS Codes

The covariates in the model will be adjusted accordingly based on the SAP for the corresponding endpoints.

1. MMRM

CCI

2. Cochran-Mantel-Haenszel test

CCI

3. ANCOVA model

CCI

4. MI

CCI

5. Proportional Odds Model

CCI

[redacted]

5.7 Appendix 7 Normal Reference Ranges for Vital Signs [9]

Table 12.1 - Heart rate by age (beats/minute) reference

Age	Awake Rate
Infant (< 1 y)	100-190
Toddler (>= 1 – < 3 y)	98-140
Preschool (>= 3 – < 6 y)	80-120
School-age (>= 6 – < 12 y)	75-118
Adolescent (>= 12 – < 15 y)	60-100
>= 15 y	60-100

Table 12.2 - Normal respiratory rate by age (breaths/minute) reference

Age	Respiratory Rate
Infant (< 1 y)	30-53
Toddler (>= 1 – < 3 y)	22-37
Preschool (>= 3 – < 6 y)	20-28
School-age (>= 6 – < 12 y)	18-25
Adolescent (>= 12 – < 15 y)	12-20
>= 15 y	12-20

Table 12.3 - Normal blood pressure by age (mm Hg) reference

Age	Systolic Pressure	Diastolic Pressure
Infant (< 1 y)	72-104	37-56
Toddler ((>= 1 – < 3 y)	86-106	42-63
Preschooler (>= 3 – < 6 y)	89-112	46-72
School-age (>= 6 – < 10 y)	97-115	57-76
Preadolescent (>= 10 – < 12 y)	102-120	61-80
Adolescent (>= 12 – < 15 y)	110-131	64-83
>= 15 y	90-120	50-80

5.8 Appendix 8 SI And US Conventional Units of Clinical Laboratory Values

SERUM CHEMISTRY	SI UNIT	CONVENTIONAL UNIT
Analyte		
Alpha Fetoprotein	IU/mL	ng/mL
Direct bilirubin	µmol/L	mg/dL
Calcium	mmol/L	mg/dL
Chloride	mmol/L	mEq/L
Creatinine	umol/L	mg/dL
Potassium	mmol/L	mEq/L
Sodium	mmol/L	mEq/L
Serum bile acid	µmol/L	µg/mL
Total bilirubin	µmol/L	mg/dL
Total cholesterol	mmol/L	mg/dL
Triglycerides	mmol/L	mg/dL
Haematology	**SI Unit**	**Conventional Unit**
Analyte		
Haematocrit	L/L	%
Haemoglobin	g/L	g/dL
Red blood count (RBC)	x10ˆ12/L	x10ˆ6/µL
Platelet count	x10ˆ9/L	x10ˆ3/µL
White blood cell count	x10ˆ9/L	x10ˆ3/µL
Fat Soluble Vitamins	**SI Unit**	**Conventional Unit**
Vitamin A	µmol/L	mg/L
Vitamin E	µmol/L	mg/L
Vitamin D (25-dihydroxy)	nmol/L	ng/mL
Urinalysis	**SI Unit**	**Conventional Unit**
Analyte		
Glucose	mmol/L	mg/dL

5.9 Appendix 9 Classification of Hepatic Function

- Categorization of Hepatic Impairment [6]

		Total Bilirubin	ALT or AST
No impairment			
Mild	A (5-6 PT)	B1: ≤ ULN B2: >1–1.5XULN	B1: > ULN B2: ANY
Moderate	B (7–9 PT)	>1.5–3X ULN	Any
Severe	C (10–15 PT)	>3X ULN	Any

- Child-Pugh Scores [10, 11]

		Point
Encephalopathy (exclusion criteria, as per Protocol A4250-012)	None	1
Ascites (exclusion criteria, as per Protocol A4250-012)	Absent	1
Bilirubin (baseline)	<2 mg/dL	1
	2-3 mg/dL	2
	>3 mg/dL	3
Albumin (baseline)	>3.5 g/dL	1
	2.8-3.5 g/dL	2
	<2.8 g/dL	3
Prothrombin time (PT) prolongation (baseline)	Less than 4 seconds above control for PT	1
	4-6 seconds above control for PT	2
	More than 6 seconds above control for PT	3
Classification 5 to 6 points : Child-Pugh A (mild liver impairment) 7 to 9 points : Child-Pugh B (moderate liver impairment) 10 to 15 points : Child-Pugh C (severe liver impairment)		

5.10 Appendix 10 PT Terms of Possible Sequelae of FSV Deficiency

CCI

[redacted]

CCI

[redacted]

CCI

[redacted]

CCI

[redacted]

CCI

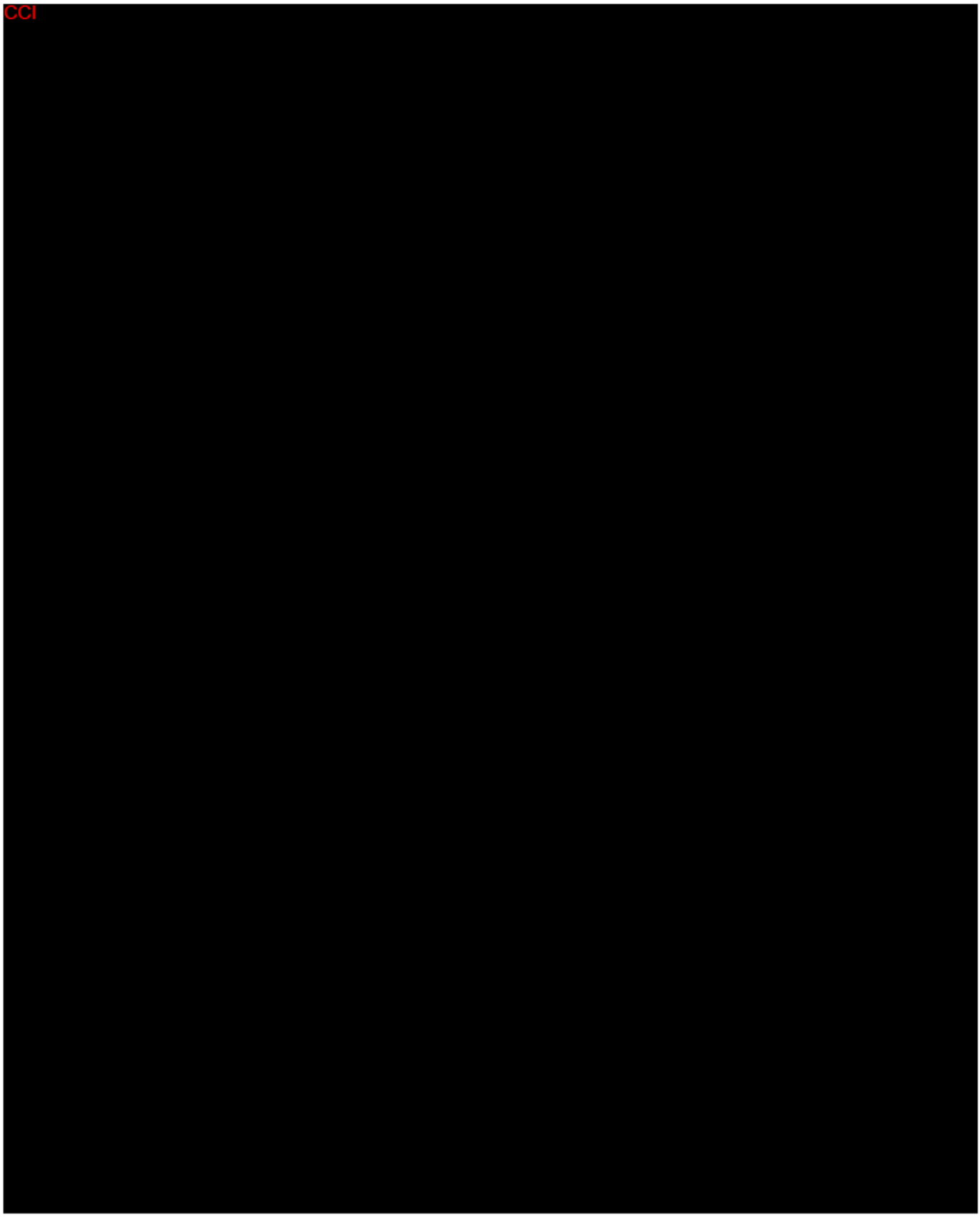
[redacted]

CCI

[redacted]

CCI

[redacted]


CCI

CCI

[redacted]

CCI

[redacted]

CCI

[redacted]

CCI

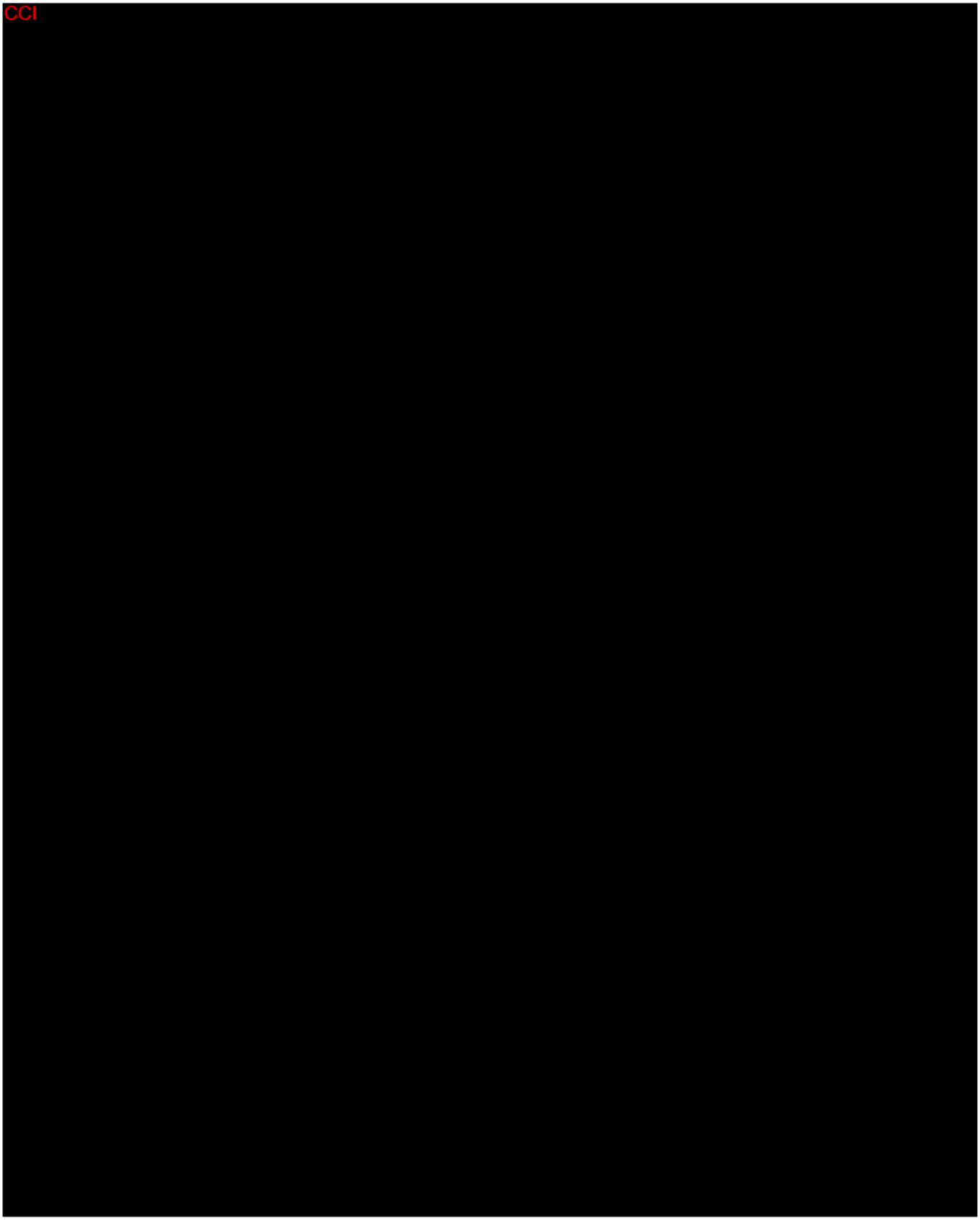
[redacted]

CCI

[redacted]

CCI

[redacted]


CCI

CCI

CCI

[redacted]

6 References

1. ICH Topic E3: Structure and Content of Clinical Study Reports (CPMP/ICH/137/95- adopted December 1995).

2. ICH Topic E9: Statistical Principles for Clinical Trials (CPMP/ICH/363/96 – adopted March 1998).

3. ICH E9 (R1): Addendum on Estimands and Sensitivity Analysis in Clinical Trials to the guideline on Statistical Principles for Clinical Trials (EMA/CHMP/ICH/436221/2017 – adopted 30 January 2020).

4. US Food & Drug Administration. Adaptive Designs for Clinical Trials of Drugs and Biologics Guidance for Industry, December 2019. (https://www.fda.gov/media/78495/download)

5. 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.

6. Aaron S. Mansfield, Michelle A. Rudek, Diana Vulih, Gary L. Smith, Pamela Jo Harris, and S. Percy Ivy (2016). The effect of hepatic impairment on outcomes in phase 1 clinical trials in cancer subjects. Clinical Cancer Research 2016 Nov 15, 22 (22): 5472-5479.

7. European Medicines Agency (EMA). Guideline on Missing Data in Confirmatory Clinical Trials (EMA/CPMP/EWP/1776/99 Rev. 1– adopted 24Jun2010)

8. US Food & Drug Administration. Adjusting for Covariates in Randomized Clinical Trials for Drugs and Biological Products. Guidance for Industry, May 2021 Revision 1. (https://www.fda.gov/media/148910/download)

9. Peds cases - Pediatrics for Medical Students - Developed by Chris Novak and Peter Gill for PedsCases.com April 21, 2016 (http://www.pedscases.com/pediatric-vital-signsreference-Chart)

10. FDA Guidance for Industry. Pharmacokinetics in Patients with Impaired Hepatic Function: Study Design, Data Analysis, and Impact on Dosing and Labeling. May 2003. Clinical Pharmacology. https://www.fda.gov/media/71311/download

11. Child Pugh Classification for Severity of Liver Disease. (https://www.merckmanuals.com/medical-calculators/ChildTurPuScore.htm)

12. CDC: https://www.cdc.gov/nccdphp/dnpao/growthcharts/resources/sas.htm

13. WHO: https://www.cdc.gov/nccdphp/dnpao/growthcharts/resources/sas-who.htm

14. Rubin, D.B., Multiple Imputation for Nonresponse in Surveys. J. Wiley & Sons, New York, 1987.