[redacted]

Statistical Analysis Plan

Study Title:	**A Phase 3, 4-Week, Parallel Group, Double Blind, Vehicle-Controlled Study of the Safety and Efficacy of ARQ-151 Cream 0.15% Administered QD in Subjects with Atopic Dermatitis**
Protocol Number and Version:	**ARQ-151-311, Original dated 31 October 2020** **Amendment 1 dated 08 June 2021** **Amendment 2 dated 18 July 2021** **Amendment 3 dated 23 May 2022** **Amendment 4 dated 28 September 2022**
Product:	**ARQ-151 Cream 0.15%**
Sponsor:	**Arcutis Biotherapeutics, Inc.** **3027 Townsgate Road, Suite 300** **Westlake Village, CA 91361**
Date:	**25 October 2022**
Version:	**Final 3.0**
Prepared by:	[redacted]

STATISTICAL ANALYSIS PLAN REVISION SUMMARY			
Version	**Version Date**	**Author**	**Summary of Changes**
Original V1.0	05-JUL-2022	[redacted]	Initial version
Final V2.0	11-Oct-2022	[redacted]	Primary estimand and multiple imputation procedure were updated as per FDA's request discussed at the Type B Pre-NDA meeting on [redacted] for a similar program. A baseline definition was added for randomized subjects who never received any study treatment (if any). Analysis day, and number of days in study derivations were updated accordingly.
Final V3.0	25-Oct-2022	[redacted]	Clarify that for subjects who discontinued early from the study due to adverse event or lack of efficacy, a subject will be considered as non-responder (for MI and non-responder imputation analyses) or missing (for observed case analyses) for all analysis visits (refer to Section 5.4) for which the subject's last dose day falls within the analysis visit window or is prior to the start of the analysis visit windows.

This statistical analysis plan will be reviewed and revised as needed. The most recent approved version will replace the previous version in place.

SIGNATURE PAGE

AUTHOR(S):

Name	Title	Signature	Date (DD-MMM-YYYY)
[redacted]	Principal Biostatistician [redacted]		

APPROVALS:

Name	Title	Signature	Date (DD-MMM-YYYY)
[redacted]	Vice President, Biometrics [redacted]		
[redacted]	Vice President, Biometrics Arcutis Biotherapeutics, Inc.		
[redacted]	Vice President, Clinical Development Arcutis Biotherapeutics, Inc.		

TABLE OF CONTENT

ABBREVIATIONS

AD	Atopic Dermatitis
AE	Adverse Event
ANCOVA	Analysis of Covariance
ATC	Anatomical Therapeutic Chemical
BMI	Body Mass Index
BSA	Body Surface Area
CDI-2	Children's Depression Inventory 2
CDLQI	Children's Dermatology Life Quality Index
CI	Confidence Interval
CMH	Cochran-Mantel-Haenszel
COVID-19	Coronavirus Disease-19
CRF	Case Report Form
CRO	contract research organization
C-SSRS	Columbia-Suicide Severity Rating Scale
CSR	Clinical Study Report
DFI	Dermatitis Family Impact
DLQI	Dermatology Life Quality Index
EASI	Eczema Area and Severity Index
ECG	Electrocardiogram
eCRF	Electronic Case Report Form
EDC	Electronic Data Capture
EM	Expectation-Maximization
ET	Early Termination
HR	Heart Rate
IP	Investigational Product
ITT	Intent to Treat
IWRS	Interactive Web Response System
LOCF	Last Observation Carried Forward
MCMC	Markov-Chain Monte-Carlo
MedDRA	Medical Dictionary for Regulatory Activities
mITT	Modified Intent to Treat
PHQ-8	Patient Health Questionnaire-8
PHQ-A	Modified PHQ-9 for Adolescents
PMM	Predictive Mean Matching
POEM	Patient-Oriented Eczema Measure
PK	Pharmacokinetic
PP	Per Protocol
PT	Preferred Term
Q1	First Quartile
Q3	Third Quartile
QD	Once Daily

SAE	Serious Adverse Event
SAP	Statistical Analysis Plan
SAS	Statistical Analysis System®
SCORAD	Scoring Atopic Dermatitis
SD	Standard Deviation
SOC	System Organ Class
TEAE	Treatment Emergent Adverse Event
TESAE	Treatment Emergent Serious Adverse Event
TLF	Tables, Listings, and Figures
WHO-DD	World Health Organization Drug Dictionary
WI-NRS	Worst Itch - Numeric Rating Score
vIGA-AD	Validated Investigator Global Assessment scale for Atopic Dermatitis

1 INTRODUCTION

This Statistical Analysis Plan (SAP) describes the planned analysis and reporting for Arcutis Biotherapeutics, Inc. clinical protocol ARQ-151-311. The analyses described in the SAP are based upon the protocol Amendment 4 dated 28 September 2022. In case of changes (note that any such changes are described in section 3.5 below) between the protocol and the SAP, the SAP will be used to guide the statistical analysis. Any deviations from the SAP will be described and justified in the final Clinical Study Report (CSR), as appropriate.

On [redacted] FDA provided the following advice in preliminary feedback prior to a pre-NDA meeting held with Arcutis for a similar program on [redacted]

> *For the primary estimand, you [Arcutis] proposed a treatment policy strategy to handle all intercurrent events, including treatment discontinuation due to adverse event or lack of efficacy. For intercurrent events of treatment discontinuation due to adverse events or lack of efficacy, we [FDA] recommend a composite strategy policy where subjects will be defined as non-responders, as we consider this to be the appropriate approach for handling such events. Your proposal to handle intercurrent events of treatment discontinuation using the treatment policy strategy can be used as part of a supportive estimand.*

The purpose of this version of the SAP is to document the change to the primary estimand and related multiple imputation strategies as requested by the FDA for a similar program.

This SAP has been developed prior to database lock, unblinding, and associated analyses. All final analyses will be performed after approval of the SAP, the clinical trial data are entered into the database, any discrepancies in the data are resolved, determination of the inclusion/exclusion of each subject from each analysis population, the database is locked, and the unblinding request form is signed.

2 STUDY OBJECTIVES AND ENDPOINTS

OBJECTIVES	ENDPOINTS
Efficacy	
To assess the efficacy of Roflumilast (also known as ARQ-151) cream 0.15% vs vehicle administered once daily	Primary efficacy endpoint:
	• Validated Investigator Global Assessment scale for Atopic Dermatitis (vIGA-AD) Success, defined as a vIGA-AD score of 'clear' (0) or 'almost clear' (1)

OBJECTIVES	ENDPOINTS
(QD) x 4 weeks to individuals 6 years of age and older with atopic dermatitis (AD).	plus at least a 2-grade improvement from Baseline at Week 4
	Secondary efficacy endpoints:
	• In subjects with a vIGA-AD score of 'Moderate' at randomization, vIGA-AD Success at Week 4 • In subjects ≥12 years old with baseline Worst Itch - Numeric Rating Score (WI-NRS) ≥ 4, achievement of at least a 4-point reduction on the WI-NRS at Week 4 • In subjects ≥12 years old with baseline WI-NRS ≥ 4, achievement of at least a 4-point reduction on the WI-NRS at Week 2 • In subjects ≥12 years old with baseline WI-NRS ≥ 4, achievement of at least a 4-point reduction on the WI-NRS at Week 1 • Achievement of at least a 75% reduction in the Eczema Area and Severity Index (EASI-75) at Week 4 • vIGA-AD of 'clear' or 'almost clear' at Week 4 • vIGA-AD Success at Week 2 • vIGA-AD Success at Week 1 • vIGA-AD of 'clear' or 'almost clear' at Week 2 • vIGA-AD of 'clear' or 'almost clear' at Week 1
	Exploratory efficacy endpoints:
	• Other continuous efficacy endpoints include change and percent change in average weekly WI-NRS and daily WI-NRS, EASI, % body surface area (BSA) affected by AD, Dermatology Life Quality Index (CDLQI/DLQI), the Dermatitis Family Impact (DFI), Scoring Atopic Dermatitis (SCORAD), and Patient-oriented Eczema Measure (POEM) at Week 1, Week 2 and Week 4. • Other categorical efficacy endpoints include EASI-50, EASI-90, EASI-100, and vIGA-AD of 'clear' at Week 4
Safety	

OBJECTIVES	ENDPOINTS
To assess the safety of Roflumilast cream 0.15% vs vehicle administered once daily (QD) x 4 weeks to individuals 6 years of age and older with AD	• Incidence and severity of adverse events (AEs) • Changes and percent change in clinical laboratory results • Changes and percent change in vital signs • The subject incidence of >5% weight loss or gain on study • Local tolerability assessments • Patient Health Questionnaire depression scale (PHQ-8) and Modified PHQ-9 for Adolescents (PHQ-A) • Children's Depression Inventory 2nd Edition (CDI-2) • Columbia-Suicide Severity Rating Scale (C-SSRS)
Pharmacokinetic	
To assess the systemic exposure of roflumilast and its N-oxide metabolite following Roflumilast cream 0.15% QD application x 4 weeks	• Plasma concentrations of roflumilast and its N-oxide metabolite

3 STUDY DESIGN

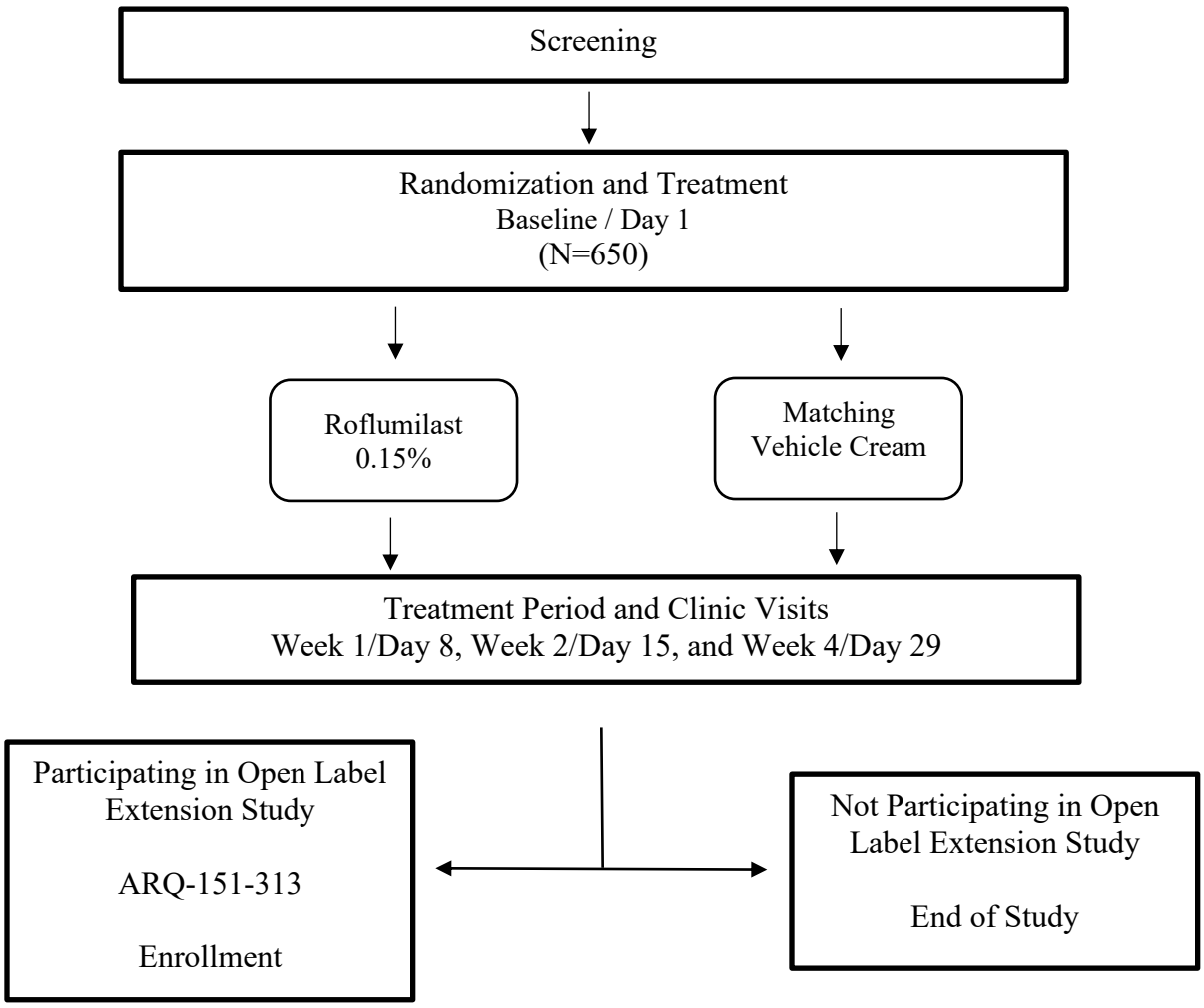
3.1 Overall Design

This is a Phase 3, parallel group, double blind, vehicle-controlled study in which Roflumilast cream 0.15% or vehicle is applied QD x 4 weeks to subjects with mild to moderate atopic dermatitis.

- Upon determination of eligibility, subjects will be randomized 2:1 to either Roflumilast cream 0.15% or matching vehicle cream. The randomization will be stratified by vIGA-AD score at baseline ('Mild' vs. 'Moderate') and by study site.
- At the Week 4 visit, subjects may be eligible to enroll in an open label extension study (ARQ-151-313) in which they will receive Roflumilast cream 0.15% QD.

The trial design is represented schematically in Figure 1.

Figure 1 **Study Schema**



A Phase 3, 4-Week, Parallel Group, Double Blind, Vehicle-Controlled Study of the Safety and Efficacy of ARQ-151 Cream 0.15% Administered QD in Subjects with Atopic Dermatitis

Approximately 650 subjects with atopic dermatitis will be randomized 2:1 to receive either:

- Roflumilast cream 0.15% or Vehicle cream

Subjects will have ≥3% BSA involvement (excluding the scalp, palms, soles) with a vIGA-AD score of '2' (mild) or '3' (moderate) for study entry

Up to 50% of the subjects will be ≥ 18 years old

3.2 Schedule of Events

Table 1 Schedule of Visits and Assessments

Study Procedure	Screen	Baseline Day 1	Wk 1 Day 8	Wk 2 Day 15	Wk 4 Day 29 / ET
Visit	**1**	**2**	**3**	**4**	**5**
Visit Window	**-30 days**	**N/A**	**+/- 3 days**	**+/- 3 days**	**+/- 3 days**
Informed consent/assent	X				
Demographics	X				
Medical and surgical history	X				
Physical examination[a]	X	X			X
I/E criteria	X	X			
Hematology, Serum Chemistries, and Urine Analysis[b]	X[b]	X[b]			X[b]
Vital signs, height, weight[c]	X	X	X	X	X
vIGA-AD, EASI, BSA, SCORAD[d]	X	X	X	X	X
WI-NRS pruritus[e]	X	X	X	X	X
POEM[f]	X	X	X	X	X
Local Tolerability Assessment[g]		X	X	X	X
CDI-2, PHQ-8, PHQ-A, C-SSRS[h]	X	X	X	X	X
DLQI, CDLQI, DFI[i]	X	X	X	X	X
Medical Photography[j]		X	X		X
Serum pregnancy test (FOCBP only)	X				
Urine pregnancy test[k]		X	X	X	X
PK draws[l]					X
Drug/vehicle application in clinic[m]		X	X	X	
Dispense/Re-dispense study medication kit[n]		X	X[o]	X[o]	X[o]
Dispense/review diary	X	X	X	X	X
Weigh study medication kit[p]		X	X	X	X
Compliance determination[q]			X	X	X

Study Procedure	Screen	Baseline Day 1	Wk 1 Day 8	Wk 2 Day 15	Wk 4 Day 29 / ET
Adverse event assessment[r]	X	X	X	X	X
Concomitant medications	X	X	X	X	X
Study Exit [s]					X

a Limited physical examination: skin (including assessment of Fitzpatrick skin type at Screening only), lungs, and heart only

[b] For all subjects entering this study under Amendment 2, to be collected at Screening, but subsequent samples will be collected only for subjects ≥12 years old (Baseline/Day 1 and Week 4/Day 29/ET). For subjects 12 to 18 years of age, if Baseline/Day 1 is within 3 weeks of Screening, the Screening results may be used.

[c] Height will be collected at Screening only. Weight should be obtained using a calibrated weight scale and the same scale should be used for a subject throughout the duration of the study. The subject should remove shoes and heavy clothing (sweaters or jackets), and empty pockets. The subject should stand with both feet in the center of the scale with their arms at their side and hold still. Record the weight to the nearest decimal fraction (for example, 25.1 kilograms). For subjects <18 years of age, measure the weight in triplicate and report the average weight in EDC. A 5% or greater weight loss (whether or not intentional or other explained) should be reported to the medical monitor.

[d] The vIGA-AD assessment will be a 5-point scale ranging from clear (0) to severe (4) and is evaluated for the entire body except the scalp, palms, and soles. EASI takes into account overall severity of erythema, infiltration/papulation, excoriation, and lichenification, in addition to extent of BSA affected. The 4 clinical signs will be graded on a 4-point scale (0 [absent] to 3 [severe]) for 4 body regions (head and neck, upper extremities, lower extremities, and trunk). Total EASI score will be calculated as a sum of scores of all 4 body regions. EASI total score will range from 0 (absent) to 72 (severe). Total BSA affected by AD will be determined for all body surfaces except the scalp, palms and soles. **The vIGA-AD assessment should be completed prior to other physician assessments**. SCORAD total score will range between 0 and 103.

[e] Subjects will self-assess their pruritus at home on a daily basis starting 7 days prior to the Baseline/Day 1 visit, and then every day thereafter. WI-NRS score will be determined by the subject assessing worst itch over the past 24 hours. The scale is from 0 (no itch) to 10 (worst itch) and this value will be recorded by the subject each day. Subjects will be trained at the Screening visit in the accurate completion of the WI-NRS. In addition, parents/caregivers of children and adolescent subjects will be trained at the Screening visit by study staff on how to assist the subject, if needed, in completing the WI-NRS.

[f] POEM will be completed by all subjects either by self or by proxy completion (for children unable to read and/or understand the POEM questionnaire, the parent/guardian/caregiver will complete the questionnaire).

[g] Local tolerability assessments should be recorded prior to study drug application for the Investigator assessment of skin irritation (Berger and Bowman skin irritation score). **Note for investigator tolerability assessments: reactions at the site of product application, which may occur post-Baseline, should be differentiated from the preexisting inflammation associated with the subject's atopic dermatitis.** The subject will assess burning/stinging (0-3 score) 10-15 minutes post drug application. **Note subject burning stinging assessment: at Day 29, subjects will provide a recall assessment of burning/stinging experienced post drug application on the previous day (Day 28).**

[h] Adolescents and adults will complete the C-SSRS (12 years of age and older). Adults will complete the PHQ-8. Adolescents (ages 12 to 17, inclusive) will complete the PHQ-A (PHQ-9 modified). Parents/caregivers will complete CDI-2 (parent report) for children 6-11 years of age, inclusive.

[i] The DLQI will be completed by subjects ≥17 years of age. The CDLQI will be completed for subjects 6 to 16 years old, inclusive. The Dermatitis Family Impact Questionnaire (DFI) will be completed by parents/caregivers for all subjects 6 to ≤17 years of age.

[j] Photography of AD lesion(s) selected by the Investigator will be performed at all investigational sites. All efforts will be made to de-identify the subjects. Canfield equipment will be used to capture photographs. Subjects who are unwilling to participate in the medical photography will be allowed to opt out of this procedure, as documented on the Informed Consent Form.

[k] A urine pregnancy test will be administered to all females of child-bearing potential. A negative result is required for continued participation in the study, and results must be available prior to dispensing of study drug at each visit.

[l] For all subjects entering this study under Amendment 2, a single PK trough draw will be collected at Day 29 only for subjects ≥12 years old. Ensure study medication was not applied in the area where PK will be drawn.

[m] Subjects to apply assigned IP during clinic visits, except for the Day 29/ET visit.

[n] It is expected that kits will be dispensed based on %BSA affected. See IP Handling Manual for details.

[o] On Day 8 and 15, dispensing of IP is optional. Site should review IP kit to ensure sufficient IP is available until the next visit and only dispense additional IP if needed. On Day 29, if the subject is unable to perform the Day 29 clinic visit due to COVID-19 restrictions (isolation, quarantine, etc.) then additional IP may need to be dispensed so IP can continue to be applied at home until the subject is able to return to the clinic to complete the Day 29 assessments (see IP Handling Manual for the process to dispense additional IP at or after Day 29).

[p] Every tube should be weighed and recorded when dispensed and returned. See IP Handling Manual for details.

[q] Compliance determination is described in the IP Handling Manual

[r] All AEs should be collected starting after the first application of the investigational product through the end of the study. All SAEs should be collected starting after the signing of the informed consent through 30 days after the last day of the application of the investigational product or the end of the study (whichever is later). Any AEs (whether serious or non-serious) and clinically abnormal laboratory test values(s) will be evaluated by the PI and treated and/or followed up for up to 30 days after end of treatment or until symptoms or value(s) return to normal, or acceptable level, as judged by the PI (if the subject is continuing into the ARQ-151-313 OLE study, then AEs from this study (ARQ-151-311) will only be followed until exit from this study).

[s] Subjects who enroll into the open label extension study (ARQ-151-313) must complete the ARQ-151-311 visit requirements at Week 4.

3.3 Treatment

Roflumilast cream 0.15% or vehicle cream will be administered QD for 28 days (+/- 3 days).

- Roflumilast cream 0.15%
- Vehicle cream

3.4 Randomization, Replacement, and Unblinding Procedures

Randomization will take place at the Baseline visit prior to first dosing. Subjects who meet all eligibility criteria will be randomized at a 2:1 ratio (drug:vehicle) to receive Roflumilast cream 0.15% QD or matching vehicle QD. The randomization will be stratified by vIGA-AD score at baseline ('Mild' vs. 'Moderate') and by study site according to a computer-generated randomization list. Kits containing tubes of study medication will be assigned to each subject using an internet-based response system (IWRS). A subject may receive more than one kit for the treatment period. The kits and tubes are blinded and each kit is numbered with a unique kit number.

The study is double-blinded, therefore neither the subjects nor the Investigator, sponsor and clinical personnel will be aware of which treatment an individual subject receives.

3.5 Changes to the Analysis from the Protocol

Rationale for change	Description of the change
A categorical summary was excluded from the SAP as it was determined that it was not necessary.	Original text: The amount of investigational product used by each subject based on tube weight will be summarized by treatment using descriptive statistics, and categorically. Changed to: The amount of investigational product used by each subject based on tube weight will be summarized by treatment using descriptive statistics.
To separate COVID patient from Per Protocol analysis set.	Original text: Per protocol (PP) population will include all subjects in the ITT population, who are at least 80% compliant with study medication application, and show no major deviations from the study protocol that would affect the interpretation of efficacy. In addition, subjects who miss the Week 4 vIGA-AD assessment specifically due to novel coronavirus disease-19 (COVID-19) disruptions will be excluded from per protocol population. Changed to: Per protocol population will include all subjects who are randomized, at least 80% compliant with study medication application, have a vIGA-AD assessment within the Week 4 visit window, and show no "major deviations" from the study protocol that would affect the interpretation of efficacy.
Added mITT population to address COVID-19 impact.	The mITT population includes all randomized subjects with the exception of subjects who missed the week 4 vIGA-AD assessment specifically due to COVID-19 disruption. This population will be used for sensitivity analysis for the primary endpoint.

4 POPULATIONS FOR ANALYSIS

4.1 Intent-to-Treat (ITT) Population

Intent-to-Treat population will include all subjects who are randomized, and all subjects will be analyzed according to the treatment group and stratum to which they were randomized.

4.2 Modified Intent-to-Treat (mITT) Population

The mITT population includes all randomized subjects with the exception of subjects who missed the week 4 vIGA-AD assessment specifically due to COVID-19 disruption. This population will be used for sensitivity analysis for the primary endpoint. All subjects will be analyzed according to the treatment group and stratum to which they were randomized.

4.3 Per Protocol (PP) Population

Per protocol population will include all subjects who are randomized, at least 80% compliant with study medication application, have a vIGA-AD assessment within the Week 4 visit window and show no "major deviations" from the study protocol that would affect the interpretation of efficacy. A complete list of major deviations from the study protocol will be created prior to unblinding and include a list of all subjects who will be excluded due to those major deviations. See section 7.2 for more details.

All subjects will be analyzed according to the actual treatment group they received and the stratum they belong to. Actual and randomized treatment will only differ if the subject received the wrong treatment throughout their participation in the study.

4.4 vIGA-AD Moderate ITT Population

vIGA-AD Moderate ITT population will be a subset of the ITT population with vIGA-AD score (randomized score) 'Moderate' at randomization.

All subjects will be analyzed according to the treatment group and the stratum to which they were randomized.

4.5 vIGA-AD Moderate PP Population

vIGA-AD Moderate PP population will be a subset of the PP population with vIGA-AD score (actual score) 'Moderate' at randomization.

All subjects will be analyzed according to the actual treatment group they received and the stratum they belong to. Actual and randomized treatment will only differ if the subject received the wrong treatment throughout their participation in the study.

4.6 WI-NRS Population

WI-NRS population will be a subset of the ITT population among subjects ≥12 years old with average weekly baseline WI-NRS score ≥ 4.

The subjects in WI-NRS population are those:

1. Completed at least 4 of 7 evaluable daily WI-NRS questionnaires during the last 7 days of the Screening period;
2. Have a mean baseline WI-NRS score ≥ 4.0, defined as the average of all non-missing scores reported during the last 7 days of the Screening period if at least 4 of 7 evaluable daily WINRS questionnaires available. If 4 or more evaluable daily questionnaires are missing, then the data will be treated as missing.

All subjects will be analyzed according to the treatment group and stratum to which they were randomized.

4.7 Safety Population

Safety population will include all subjects who are enrolled and received at least one confirmed dose of study medication.

Subjects will be analyzed based on the treatment group received and the stratum they belong to. Actual and randomized treatment will only differ if the subject received the wrong treatment throughout their participation in the study.

4.8 Pharmacokinetic (PK) Population

Pharmacokinetic population will include all subjects receiving the active drug with quantifiable plasma concentrations of roflumilast.

5 GENERAL CONSIDERATIONS

Formats and layouts of tables, listings, and figures (TLF) will be provided in a separate document (output general layout is described in Appendices

Appendix **1**).

5.1 Sample Size

There are approximately 650 subjects planned for this study. In order to have the desired power for the secondary endpoint of vIGA-AD Success in subjects with a vIGA-AD score of 'Moderate' at randomization, approximately 490 of the subjects to be accrued will have vIGA-AD score of 'Moderate' at randomization. Randomization will be stratified by vIGA-AD score ('Mild' vs. 'Moderate') and by study site.

The sample size of 650 subjects provides approximately 95% power to detect an overall 15% difference (Odds Ratio = 2.1) between treatment groups on vIGA-AD success at Week 4 at α=0.05 using a 2-sided stratified Cochran-Mantel-Haenszel (CMH) test. The results from a recent Phase 2 study (ARQ-151-212) of Roflumilast cream 0.15% compared to vehicle treatment were used to estimate the treatment difference. Specifically, in the Phase 2 trial, approximately 37% of subjects demonstrated vIGA-AD Success at Week 4 in the ARQ-151 0.15% group compared to 22% in the vehicle group. The sample size of 490 also provides approximately 90% power to detect an overall 17% difference (Odds Ratio = 2.1) between treatment groups (28% of vIGA-AD Success at Week 4 in vehicle treatment) on vIGA-AD Success at Week 4 among subjects with vIGA-AD score 'moderate' at randomization. The same testing method, the stratified CMH test, will be used as for the primary endpoint.

To control for familywise type I error at level of 0.05, the secondary endpoint of vIGA-AD Success at Week 4 in subjects with vIGA-AD of 'Moderate' at randomization will only be tested if the primary endpoint demonstrates statistical significance. In addition, the remaining secondary endpoints will be inferentially tested only if the primary and secondary endpoint (vIGA-AD Success at Week 4 for subjects with vIGA-AD score of 'Moderate' at randomization) comparisons are statistically significant using the hierarchical testing procedure by partitioning of alpha (see Section 6.5 for more details).

5.2 Baseline

Unless otherwise specified, baseline value will be defined as the last non-missing assessment prior to or concurrently with the first study treatment dosing* (including unscheduled/retest assessments). If the last non-missing assessment is performed on the same date as the first study treatment administration* and time is not available, it is assumed that the assessment took place prior to IP application*, per study site training, and the assessment will be considered as baseline, except for adverse events (AEs) and medications starting on the first study treatment dose administration date which will be considered postbaseline.

Average weekly baseline WI-NRS is defined as the average of all non-missing scores reported during the last 7 days prior to treatment* if at least 4 of 7 evaluable daily WI-NRS questionnaires

available. Daily baseline WI-NRS is defined as the last non-missing assessment prior to or concurrently with the first study treatment dosing*. Day 1 WI-NRS score will be used to calculate the value for week 1 only when it is collected after the application of the first study drug*. If Day 1 WI-NRS score is collected prior to or concurrently with the application of the first study drug*, then the Day 1 WI-NRS score will be included in baseline calculation.

For investigator/subject tolerability assessments, baseline is derived as the measurement taken on the day of first application of study drug*.

* For randomized subjects who discontinued study before the first application of study treatment, the date of randomization will be considered instead of the date of the first application of study treatment.

5.3 Reference Start Date and Analysis Day

Analysis day will be calculated from the first study treatment administration date* and will be used to derive start/end day of assessments or events.

Analysis day = (Date of event – Date of first dose administration*) + 1 if date of event is on or after the date of first dose administration of study treatment*;

= (Date of event – Date of first dose administration*) if date of event is before the date of first dose administration of study treatment*.

In the situation where the assessment/event date is partial or missing, analysis day will be missing.

* For randomized subjects who discontinued study before the first application of study treatment, the date of randomization will be considered instead of the date of the first application of study treatment.

5.4 Windowing Conventions

Visits will be analysed as scheduled. Unscheduled, early termination visits, and/or retest measurements will only be included if a scheduled measurement is not available and the early termination or unscheduled/retest measurement falls within the analysis visit windows as described in Table 2, Table 3, and Table 4 when appropriate. Unscheduled/retest measurements will be listed.

If there is more than one assessment for a given timepoint and analysis visit when a scheduled measurement is not available, the assessment closest to the target day will be considered.

Table 2 Analysis Visit Windows for Efficacy Endpoints, Vital Signs, CDI-2, PHQ-8, PHQ-A, C-SSRS and Local Tolerability Assessment

Analysis Visit	Target Day	Lower Limit	Upper Limit
Week 1	8	2	11
Week 2	15	12	22
Week 4	29	23	43

Table 3 Analysis Visit Windows for Clinical Laboratory, Physical Examination and PK assessment

Analysis Visit	Target Day	Lower Limit	Upper Limit
Week 4	29	23	43

Table 4 Windows for the derivation of Average Weekly WI-NRS

Days for calculation of weekly average	Week (Derived)
(-7, -1)*	Baseline
(1, 7)*	Week 1
(8, 14)	Week 2
(15, 21)	Week 3
(22, 28)	Week 4

* Day 1 WI-NRS score will be used to calculate the value for week 1 only when it is collected after the application of the first study drug (randomization for randomized subjects who were never treated with study drug). If Day 1 WI-NRS score is collected prior to or concurrently with the application of the first study drug (randomization for randomized subjects who were never treated with study drug), then the Day 1 WI-NRS score will be included in baseline calculation.

Note: With the caveat described in footnote '*', if more than one WI-NRS score is available on the same day, the worst score of the day will be considered in the analyses.

5.5 Derived Variables

All questionnaire scores will be derived by Biostatistics in the ADaM datasets using the formulas defined below, even if calculated scores are present in the EDC database. All pre-calculated scores will be ignored for analysis.

- With the following exception, vIGA-AD Success = vIGA-AD of 'Clear' (0) or 'Almost Clear' (1) plus at least a 2-grade improvement from Baseline. The exception is that for subjects who discontinued early from study due to an AE or lack of efficacy, the subject will be considered as not having a vIGA-AD success (for MI and non-responder imputation analyses) or missing (for observed case analyses) for all pre-specified analysis visits (refer to Section 5.4) for which the subject's last dose day falls within analysis visit window or is prior to the start of the analysis visit window.

- Average weekly WI-NRS = Average weekly WI-NRS pruritus score will be calculated as the sum of the daily WI-NRS scores reported during a specific week (in a 7-day period; refer to Section 5.4) of the study divided by the number of days with non-missing scores for that week. A minimum of 4 days of observations are needed to calculate an average weekly WI-NRS pruritus score. Otherwise, the corresponding average weekly WI-NRS pruritus score will be considered missing.

- With the following exception, a WI-NRS 4-point reduction = achievement of a 4- point reduction in average weekly WI-NRS pruritus score compared to average weekly WI-NRS baseline, calculated only for the subjects ≥12 years old with average weekly WI-NRS score of ≥ 4 at baseline. The exception is that for subjects who discontinued early from study due to an AE or lack of efficacy, the subject will be considered as not having a WI-NRS 4-point reduction (for MI and non-responder imputation analyses) or missing (for observed case analyses) for all pre-specified analysis visits (refer to Section 5.4) for which the subject's last dose day falls within the analysis visit window or is prior to the start of the analysis visit window.

- EASI total score = $0.1\ (E_h + I_h + E_{Xh} + L_h)\ A_h + 0.2\ (E_u + I_u + E_{Xu} + L_u)\ A_u + 0.3\ (E_t + I_t + E_{Xt} + L_t)\ A_t + 0.4\ (E_l + I_l + E_{Xl} + L_l)\ A_l$
  for subjects '≥ 8 years old
  and
  $0.2\ (E_h + I_h + E_{Xh} + L_h)\ A_h + 0.2\ (E_u + I_u + E_{Xu} + L_u)\ A_u + 0.3\ (E_t + I_t + E_{Xt} + L_t)\ A_t + 0.3\ (E_l + I_l + E_{Xl} + L_l)\ A_l$
  for subjects < 8 years old

  where E, I, Ex, L, and A denote erythema, induration, excoriation, lichenification, and area, respectively, and h, u, t, and l denote head, upper extremities, trunk, and lower extremities, respectively. Scalp, palms, and soles may be treated with investigational product in this study but will be excluded from the EASI assessment. If a subject turns 8 years old during the study, the formula used at Screening will continue to be used through the study duration of the subject's participation in the study.

- With the following exception, EASI-50, EASI-75, EASI-90, and EASI-100 = Achievement of at least a 50%, 75%, 90%, or 100% reduction from baseline in EASI total score, respectively. The exception is that for subjects who discontinued early from study due to an AE or lack of efficacy, the subject will be considered as not having reached EASI-50, EASI-75, EASI-90 and EASI-100 (for MI and non-responder imputation analyses) or missing (for observed case analyses) for all pre-specified analysis visits (refer to Section 5.4) for which the subject's last dose day falls within the analysis visit window or is prior to the start of the analysis visit window.

- DLQI Score = sum of the 10 questions (individual questions scored as Very much=3, A lot=2, A little=1, Not at all=0, Not relevant=0, Question 7: Yes=3, if No, then follow the same score as A lot, A little, Not at all), ranging from 0 to 30. If 1 item is missing, it is scored as 0 for that item. If 2 or more items are missing, the score should not be calculated.

- CDLQI Score = sum of the 10 questions (individual questions scored as Very much=3, Quite a lot=2, Only a little=1, Not at all=0; Question 7: if the last week was school time, the question was scored as Very much=3, Quite a lot=2, Only a little=1, Not at all=0, with Prevented school recoded to 3, and if the last week was holiday time, the standard responses apply), ranging from 0 to 30. If 1 item is missing, that item is scored as 0. If 2 or more items are missing, the score should not be calculated.

- DFI Score = sum of the 10 questions (individual questions scored as Very much=3, A lot=2, A little=1; Not at all=0), ranging from 0 to 30. If 1 item is missing, it is scored as 0 for that item. If 2 or more items are missing, the score should not be calculated.

- PHQ-8 = sum of the 8 questions (individual questions scored as Not at all=0, Several days=1, More than half the days=2, and Nearly every day=3), ranging from 0 to 24. If more than 1 item is missing the score should not be calculated. If 1 item is missing the score is calculated as (sum of answered items*8)/number of answered items (=7).

- Modified PHQ-A = sum of the 8 questions (individual questions scored as Not at all=0, Several days=1, More than half the days=2, and Nearly every day=3), ranging from 0 to 24. If more than 1 item is missing the score should not be calculated. If 1 item is missing, the score is calculated as (sum of answered items*8)/number of answered items (=7).

- CDI-2 total score is a sum of the 17 questions based on the scoring grid (individual questions scored as much or most of the time=0, often=1, some of the time=2, Not at all=3 for questions Q2, Q7, Q13, Q14, and Q16; individual questions scored as much or most of the time=3, often=2, some of the time=1, Not at all=0 for all other questions), ranging from

0 to 51. All CDI-2 raw scores can be prorated using the following formula, rounding to the nearest whole number:

- Prorated score = (Obtained raw score for scale) * (Total # of items on scale/subscale) / Total # items on scale/subscale with responses
- CDI-2 emotional problem scale is a sum of 9 questions (Q1, Q3-6, Q8, Q10-12). If more than 1 item is missing the score should not be calculated. If 1 item is missing the score is calculated as (sum of answered items*9)/number of answered items (=8).
- CDI-2 functional problem scale is a sum of 8 questions (Q2, Q7, Q9, Q13-17). If more than 1 item is missing the score should not be calculated. If 1 item is missing the score is calculated as (sum of answered items*8)/number of answered items (=7).
- CDI-2 total score is a sum of 17 questions. If more than 2 items are missing the score should not be calculated. If 1 item is missing the score is calculated as (sum of answered items*17)/number of answered items (=16). If 2 items are missing the score is calculated as (sum of answered items*17)/number of answered items (=15).

- POEM = sum of the 7 questions (individual questions scored as No days = 0, 1 to 2 days = 1, 3 to 4 days = 2, 5 to 6 days = 3, Every day = 4), ranging from 0 to 28. If 1 question is left unanswered this is scored 0 and the scores are summarized and expressed as usual out of a maximum of 28. If 2 or more questions are left unanswered the questionnaire is not scored.
- SCORAD = [Overall BSA affected by AD / 5] + [Intensity score*7/2] + subjective symptoms score (pruritus + sleep loss); SCORAD score will be set to missing if information for any of the three measures is missing.

5.6 Descriptive Statistics

All continuous variables will be summarized by presenting the number of subjects, mean, standard deviation (SD), median, Q1, Q3, minimum, and maximum. For PK endpoints, geometric statistics including geometric mean and coefficient of variation (CV) will also be provided.

Categorical variables will be presented as frequencies and percentages.

Summary tables will be presented by visit, when applicable.

Change from baseline will be calculated as:

Assessment value at postbaseline visit X – baseline value.

Percent change from baseline will be calculated as:

(Assessment value at postbaseline visit X – baseline value) × 100% / (baseline value)

Percent change from baseline will be missing in situation where baseline value equals to 0.

5.7 Statistical Tests

Unless otherwise specified, all statistical tests will be two-sided and will be performed with a significant level of 0.05. Confidence intervals (CIs) will be two-sided with 95% coverage.

5.8 Handling of Retests, Unscheduled Visits, and Early Termination Data

Retests measurements, Unscheduled measurements, and ET visit assessments will be included in analysis and be summarized via analysis visit windowing according to the windowing conventions in section 5.4.

All data from retest, unscheduled measurements and ET visit assessments will be listed.

5.9 Software Version

All analyses will be performed using SAS® software Version 9.4 or higher.

6 STATISTICAL CONSIDERATIONS

6.1 Adjustments for Covariates

Covariates for this study include pooled study site and baseline vIGA-AD (vIGA-AD=2 - Mild vs. vIGA-AD=3 - Moderate at randomization). Subgroup analyses will be generated for the baseline covariates.

6.2 Handling of Dropouts or Missing data

See Appendix 2 for handling of completely or partially missing dates for prior and concomitant medications and AEs.

Unless otherwise specified, missing safety data will not be imputed.

6.2.1 Multiple Imputation

All subjects, regardless of completion status, will have available data assigned to the pre-specified analysis visit using the analysis windows defined in Section 5.4, including the last available data

of subjects who prematurely withdraws from the study. To comply with the definition of the primary estimand (Section 12.1.1), for subjects who discontinue due to lack of efficacy or adverse event, efficacy data assigned to a pre-specified analysis visit will be removed from the source data used for the multiple imputation process if subject's last dose day falls within the analysis visit window or is prior to the start of the analysis visit window used to assign data to a pre-specified analysis visit. Similarly, WI-NRS weekly averages will be removed from the source data used for multiple imputation if subject's last dose day falls within the analysis visit window or is prior to the start of the interval used to compute the weekly average as defined in Section 5.4. This procedure will ensure that the data collected on or after intercurrent events are not used in the imputation process.

For the primary efficacy endpoint of vIGA-AD success at Week 4 and the secondary endpoint of vIGA-AD success at Week 4 among subjects with a 'Moderate' randomized vIGA-AD score, the primary analysis will impute missing values using a Predictive Mean Matching (PMM) sequential-regression multiple imputation model for the ITT population. This is a three-step process.

1. The first step is to understand the pattern of missingness. In order to perform the multiple imputation, a monotone missing pattern has to be achieved. For example, if there exist values for baseline and Week 4 visits, but missing values for the Week 1 or 2 visits, the Markov-Chain Monte-Carlo (MCMC) method will be used to impute the small amount of missing data that may be missing at the intermediate visits that is required to make the missing data pattern monotone before applying the multiple imputation algorithm. This method uses a non-informative Jeffreys prior to derive the posterior mode from the expectation-maximization (EM) algorithm as the starting values for the MCMC method. The MCMC method will use the seed 6457149. The vIGA-AD score will be treated as a continuous variable for this step and the model will include the vIGA-AD scores at baseline, Week 1, Week 2, and Week 4. To avoid values that could not be observed in practice, imputed values will be rounded to the nearest integer (Round=1 option in PROC MI) in the range of 0 to 4.

   To determine the number of multiply-imputed datasets to be created at this step, the proportion of datapoints with non-monotone pattern across all visits and subjects will first be derived as follows:

$$\frac{number\ of\ non\ monotone\ visits\ accross\ all\ visits\ and\ subjects}{total\ number\ of\ expected\ visits\ across\ all\ subjects} * 100$$

   Then, the following table will be used:

Non-monotone Missing Data	Number of Imputed Datasets
≤ 2%	1
> 2% to ≤ 5%	3
> 5%	10

2. Once the monotone pattern is achieved, the next step is to implement the imputation algorithm. For this, the PMM regression method will be used. This method is particularly helpful if the normality assumption is violated. For subjects with complete data up to a particular visit, a PMM regression model will be fit that includes the outcome at that visit as the dependent variable and as independent variables, the treatment group, pooled study site, and vIGA-AD score outcomes at previous visits, using a seed of 482371. For other scales/questionnaires, the actual baseline vIGA score will also be included as an independent variable. This process will be repeated 25 times, resulting in a total of 25 to 250 complete analysis datasets, depending on the number of imputed monotone datasets that are required. The seed may be changed after unblinding in case of any issues with the imputation process, and it will be documented in the CSR if any change is required.

3. For each of the 25 to 250 completed dataset, the necessary derived variable will be computed as defined in Section 5.5 and analyzed using a CMH analysis, adjusted for the pooled study site and randomized vIGA-AD score for the primary efficacy endpoint and adjusted for the pooled site for the secondary efficacy endpoint. Results will be combined into one multiple imputation inference as follows:
    a. Common proportion of success, common Mantel-Haenszel (MH) proportion difference (and associated 95% CI), and common MH odds ratio (and associated common 95% CI) will be combined using PROC MIANALYZE based on Rubin's rule. Common MH odds ratios and associated common 95% CI will first be normalized using a log-transformation before being combined using PROC MIANALYZE. The resulting combined common MH odds ratio and associated combined common 95% CI will be back-transformed to the arithmetic scale before being presented in a table.
    b. For the combined common proportion of success, the associated combined common 95% CI will be calculated as per Lott and Reiter multiple imputation Wilson interval method[1].
    c. Two p-values will be produced for each analysis:
        i. The primary p-value will be obtained from a multiple imputation CMH test, where CMH general association statistics and their standard errors obtained

from the analysis of each multiply-imputed dataset will be transformed as per Wilson-Hilferty[2] before being combined using PROC MIANALYZE based on Rubin's rule. Because the Wilson-Hilferty transformation is a monotone transformation, the p-value for the CMH test is the one-sided p-value from the t distribution. This p-value will be the primary p-value used to evaluate the result according to the multiple testing strategy described in Section 6.5.

i. Should the common MH odds ratio from the analysis of at least one of the multiply-imputed datasets be not estimable, combined common MH odds ratio and associated combined common 95% CI and p-value will not be presented. Under such circumstance, conclusions will be based on the p-value obtained from a multiple imputation test of the proportion difference, where the common proportion difference using MH weights and associated common standard errors based on the Sato variance estimator obtained from the analysis of each multiply-imputed dataset will be combined using PROC MIANALYZE based on Rubin's rule.

Similar multiple imputation method will be used for the average weekly WI-NRS and EASI total score.

For the average weekly WI-NRS (refer to Section 5.5), it's the missing average weekly WI-NRS data (i.e., those that cannot be computed because only 3 or less non-missing assessments are available during a given week) at Weeks 1, 2, 3, and 4 that will be multiply imputed, not the missing daily WI-NRS data. Since the missing average weekly WI-NRS values have a precision of 1 decimal place, the MCMC imputation for the non-monotone missing data will be restricted to values between 0 and 10, rounded to the 1st decimal (i.e., 0.1). Imputation of missing data for WI-NRS will be based on the ITT population, i.e., all randomized subjects will be included in the imputation process.

For the EASI total score (refer to Section 5.5), it's the missing EASI total score data at Weeks 1, 2, and 4 that will be imputed, not the missing EASI question score data. Since the missing EASI total scores have a precision of 1 decimal place, the MCMC imputation for the non-monotone missing data will be restricted to values between 0 and 72, rounded to the 1st decimal (i.e., 0.1). Imputation of missing data for EASI total score will be based on the ITT population.

6.2.2 Non-responder Imputation Analysis

If assessment of vIGA-AD after baseline is missing, the subject will be considered as non-responder (for example, no vIGA-AD success in the analysis of vIGA-AD success). For the vIGA-

AD success endpoints, subjects with missing baseline vIGA-AD score will be considered as non-responder for these endpoints only. That is, for the secondary efficacy endpoints of vIGA-AD score of clear (0) or almost clear (1) and vIGA-AD score of clear (0), the post-baseline assessment at a specific visit will be established based on the vIGA-AD score at that visit only, regardless of the availability of the baseline vIGA-AD score.

Similar imputation method will be used for 4-point reduction on the average weekly WI-NRS and EASI-75. For EASI-75, subjects with missing baseline EASI total score will be considered as non-responder. For 4-point reduction on the average weekly WI-NRS, subjects with missing baseline average weekly WI-NRS will not be imputed since the analysis of WI-NRS endpoints will be performed based on the WI-NRS population from which subjects with a missing baseline average weekly WI-NRS are excluded.

6.2.3 Tipping Point Analysis

As a sensitivity analysis to the multiple imputation analysis as described in Section 6.2.1 for the vIGA-AD success primary endpoint and secondary endpoint of vIGA-AD success among subjects with Moderate disease at baseline, a tipping point analysis will be performed in order to determine the inflection point at which the inference under the missing not at random (MNAR) assumption changes substantially.

The sensitivity analysis for the primary endpoint will be performed by using a specified sequence of shift parameters. The range of shift parameters to be included in this analysis are 0 to 2 by 0.2 for active and -2 to 0 by 0.2 for Vehicle. The values at which the results of the primary analysis are shifted from significant (i.e., $\alpha \leq 0.05$) to non-significant (i.e., $\alpha > 0.05$) will be determined. Steps 1 and 3 of the analysis will be the same as for the multiple imputation analysis as described in Section 6.2.1. However, Step 2 of the analysis is where the shift parameters will be applied.

Imputed values for subjects who discontinue due to lack of efficacy or adverse event will be handled as described in Step 2 of Section 6.2.1 to ensure that these subjects are analyzed as non-responders at all visits on or after discontinuation of treatment.

6.3 Interim Analysis

No interim analysis is planned for this study.

6.4 Multicenter Studies

The study will be conducted at approximately 60 study sites in the US, Canada, and Poland. During the conduct of the study, additional countries and/or sites may be added if necessary.

Sites with less than 16 subjects in the ITT population will be pooled within a country:

a. Sites with less than 16 subjects will be ordered from lowest to highest in terms of number of ITT subjects. In case of ties, the ordering for tied sites will be determined according to the site identification number (from smallest to largest).
b. Sites will be combined beginning at the smallest until the resulting pooled site contains at least 16 ITT subjects with at least 1 subject in each treatment group. The sites pooled in this way will be considered as a single site in the statistical analyses.
c. The process described above will resume for the remaining sites not meeting the criterion of at least 16 ITT subjects with at least 1 subject in each treatment group. If the final set of pooled sites does not meet the criterion of at least 16 ITT subjects with at least 1 subject in each treatment group, the final set will be pooled with the preceding pooled site.
d. If there is only one site with less than 16 ITT subjects with at least 1 subject in each treatment group, then this site will be combined with the site with the second lowest number of subjects. As above, in the case of ties, the ordering for tied sites will be determined according to the site identification number (from smallest to largest).

As a sensitivity analysis of different pooling strategy, sites that have randomized less than 50% of the number of randomized subjects at the site with the largest number of randomized subjects, those sites will be pooled within each country. The sensitivity analysis of pooling strategy will be applied for the primary endpoint only.

6.5 Multiple Comparisons/Multiplicity

To control for familywise type I error at level of 0.05, the secondary endpoint of vIGA-AD Success at Week 4 for subjects with vIGA-AD score of 'Moderate' at randomization will only be tested if the primary endpoint demonstrates statistical significance. In addition, the remaining secondary endpoints will be inferentially tested only if the primary and the secondary endpoint (vIGA-AD Success at Week 4 for subjects with vIGA-AD score of 'Moderate' at randomization) comparisons are statistically significant using hierarchical testing procedure by partitioning of alpha.

Upon successful demonstration of statistical significance for the primary and above secondary endpoint, the remaining endpoints will be grouped into secondary endpoint family 1, comprised of the 4-point reduction on the WI-NRS endpoint, at Week 4, Week 2 and Week 1, and secondary endpoint family 2, comprised of the endpoints of EASI-75 at Week 4, vIGA-AD of 'clear' or 'almost clear' at Week 4, vIGA-AD of success at Week 2 and Week 1, vIGA-AD of 'clear' or 'almost clear' at Week 2 and Week 1. An alpha level of 0.03 will be used to test the endpoints in the secondary endpoint family 1 sequentially. An alpha level of 0.02 will be used to test the endpoints in secondary endpoint family 2 sequentially.

In addition to the partitioning of the overall 0.05 alpha into two families, the Fallback Method will be applied. The fallback method is a modification of the fixed-sequence method, providing opportunity to test an endpoint later in the sequence even if an endpoint tested early in the sequence has failed to show statistical significance. The order of the endpoints remains important. The appeal of the fallback method is that if an endpoint later in the sequence has a robust treatment effect while the preceding endpoint is unsuccessful, there is a modest amount of alpha retained as a fallback to allow interpretation of that endpoint without inflating the Type I error rate. Applying the fallback method begins by dividing the total alpha (not necessarily equally) among the endpoints and maintains a fixed sequence for the testing. In this study, the Fallback Method will be applied to the fixed sequence of testing Family 1, and then Family 2.

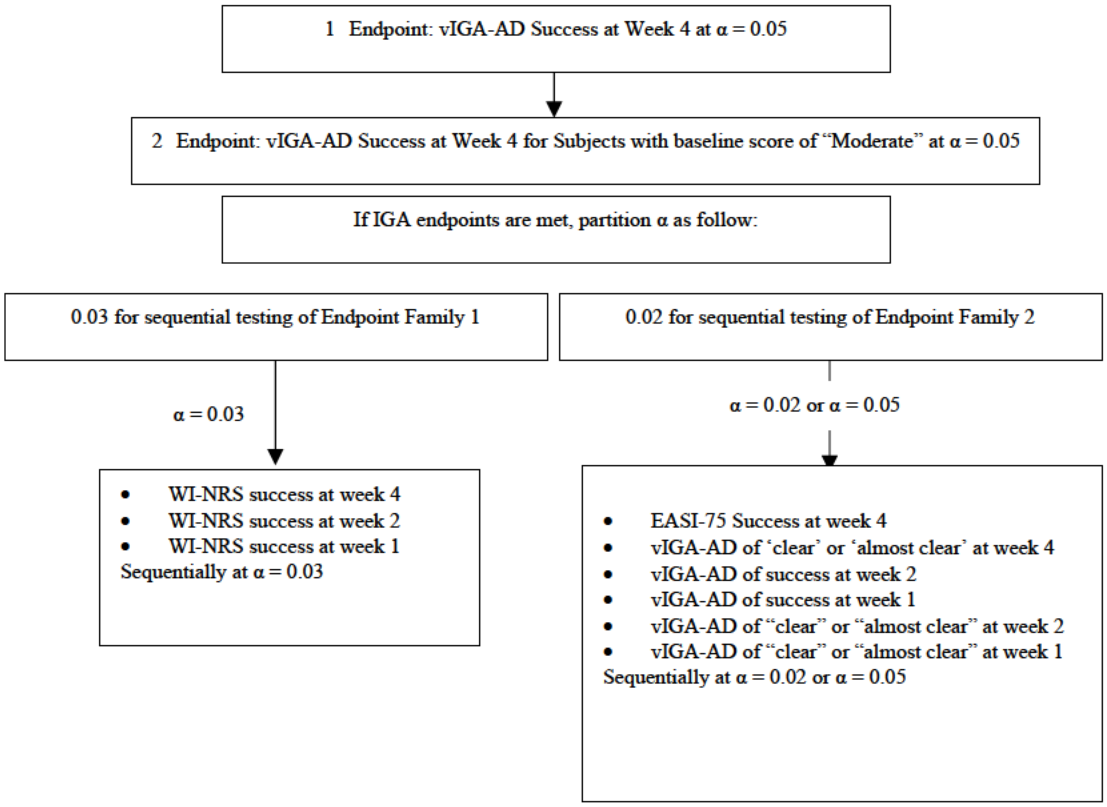
As the testing sequence progresses, a successful test preserves its assigned alpha as "saved" ("unused" or "accumulated") alpha that is passed along to the next test in the sequence, as is the case for the sequential method. This accumulated alpha is added to the prospectively assigned alpha (if any) of that next endpoint and the summed alpha is used for testing that endpoint. Thus, as sequential tests are successful, the alpha accumulates for the endpoints later in the sequence; these endpoints are then tested with progressively larger alphas.

In this study, the Fallback Method will be applied following this sequence:

Family 1: Testing will proceed at the 0.03 level sequentially within Family 1. Should all 3 endpoints in Family 1 be statistically significant at the 0.03 level, then the full 0.03 alpha will be carried to Family 2. Family 2 would then be tested at the full (α=0.02+0.03=0.05).

Should, anywhere during the sequential testing of Family 1, there is a p-value >0.03, the testing within Family 1 will stop, and no additional alpha can be carried over to Family 2.

Figure 2 Multiple Testing Scheme



Achievement of vIGA-AD success is a score of "clear" or "almost clear" plus a 2-grade improvement from baseline.

WI-WRS Success is a 4-point reduction in WI-NRS among subjects ≥12 years old with WI-NRS ≥ 4 at baseline.

EASI-75: achievement of at least a 75% reduction in the Eczema Area and Severity Index

6.6 Examination of Subgroups

Subset analysis for the following subgroups will be performed for the primary and secondary efficacy endpoints:

- Age group (6 – 11 years vs. 12 - 17 years vs. ≥18 years),
- Gender (male vs. female),
- Race (White vs. Black or African American vs. Asian vs. other),
- Ethnicity (Hispanic vs. Non-Hispanic),
- Randomized vIGA-AD score (mild (2) vs. moderate (3)),

- Actual baseline vIGA-AD score (mild (2) vs. moderate (3)),
- Baseline % BSA (<10% vs. ≥ 10%),
- %BSA categories based on tertiles,
- Baseline EASI total score (≤7 vs >7),
- Baseline EASI total score based on tertiles
- Fitzpatrick skin type at Screening (Type I, II and III vs. Type IV, V, and VI),
- Prior inadequate response, intolerance, or contraindication to Topical Corticosteroids (yes vs. no),
- Facial Involvement (yes vs. no)

For subgroup based on tertiles, tertiles will be derived using pooled data from both treatment group based on the ITT population.

Details on these analyses are described in Section 12.4.

7 STUDY SUBJECTS

7.1 Disposition of Subjects

All subjects who provide informed consent will be accounted for in this study. The number of subjects who were screened and who failed screening (screen failures) will be presented. The reasons for screen failure will be presented for all screened subjects who failed screening.

The number of subjects randomized will be presented by treatment group. The number and percentage of the subjects included in each analysis population will be provided by treatment group. The number and percentage of the subjects who completed the study, who discontinued the study, the reasons for study discontinuation, and early termination due to COVID-19 disruption will be presented by treatment group. The percentages will be calculated using the number of the randomized subjects as denominator.

Number of days in the study will be summarized with descriptive statistics by treatment group and overall. For each subject, the number of days in the study will be calculated as following:

Number of days in study = Date of completion/discontinuation – 1st dose date* +1

* For randomized subjects who discontinued study before the first application of study treatment, the date of randomization will be considered instead of the date of the first application of study treatment.

A listing of subject's disposition and randomization will be provided. Information on first screening for subjects who were rescreened, including the rescreened subject identifier and the reason for first screening failure, will be presented under the first screening subject identifier. The reason for screening failure will be listed as well.

A table of randomized strata vs. actual strata will be provided if there is any mis-randomization discrepancy.

7.2 Protocol Deviations

A data review will be conducted before database lock by the Medical Monitor and the Sponsor to classify protocol deviations as minor or major.

The number and percentage of subjects with at least one important protocol deviation (including important protocol deviations associated with COVID-19) will be summarized by deviation category and treatment group using the safety analysis set.

A listing of all protocol deviations will also be provided. The protocol deviations associated with COVID-19 and major PDs will be flagged.

8 DEMOGRAPHIC AND OTHER BASELINE CHARACTERISTICS

Demographics and baseline characteristics will be summarized with descriptive statistics using the ITT and safety population. The list of demographics and baseline characteristics to be summarized will include:

- Age (years)
- Age Group: 6-11, 12-17, 6-17 and ≥18, 18-64 and ≥65
- Childbearing potential
- Sex at birth
- Ethnicity
- Race*
- Baseline Height (cm)

- Baseline Weight (kg)
- Baseline Body Mass Index (BMI) (kg/m^2): Adult, Child/Adolescent
- Fitzpatrick Skin Type
- Prior failure of Topical Corticosteroids, Topical Calcineurin Inhibitors, Eucrisa
- Atopic Dermatitis involvement on the face, on the eyelids
- Baseline vIGA-AD
- Average weekly baseline WI-NRS
- Daily baseline WI-NRS
- Baseline BSA (%)
- Baseline BSA (%) Group - <10% and ≥10%, tertile groups
- Baseline EASI total score
- Baseline EASI score group - ≤7 and >7, tertile groups
- Baseline SCORAD
- Baseline DLQI/CDLQI
- Baseline DFI
- Baseline POEM
- Baseline PHQ-8
- Baseline PHQ-A
- Baseline CDI-2

*Subjects who reported more than one race will be summarized as 'Multiple' races in the table. All races selected will be displayed in the listing.

Adult: BMI (kg/m^2) = (weight in kg)/ [(height in cm/100)^2]. Baseline height will be used to derive BMI for each visit since height is not collected at all visits.

Child and Adolescent (6 to 17 years): After BMI is calculated using the same formula above for children and teens, it is expressed as a percentile obtained from either a graph or a percentile calculator. Percentiles will be calculated using data files and instructions provided by the CDC (https://www.cdc.gov/growthcharts/percentile_data_files.htm). Baseline height will be used to derive BMI for each visit since height is not collected at all visits.

A listing of all demographics, analysis population flag, reason not included in the efficacy analysis will be provided.

9 SURGICAL AND MEDICAL HISTORY

Surgical and medical history will be coded according to the Medical Dictionary for Regulatory Activities (MedDRA), Version 25.0.

Surgical and medical history will be summarized by system organ class (SOC) and preferred term (PT) using the safety analysis set. A subject who experienced the same surgical and medical history event multiple times will be counted only once for the corresponding PT. Similarly, if a subject experienced multiple surgical and medical history events within the same SOC, the subject will be counted only once for that SOC. Surgical and medical history events will be sorted alphabetically by SOC and within each SOC the PT will be presented by descending frequency in the safety analysis set.

10 PRIOR AND CONCOMITANT MEDICATIONS

Medications will be coded according to the World Health Organization Drug Dictionary (WHO-DD) B3 September 2022.

Prior medications are defined as any medication started and discontinued prior to the first study treatment dosing. Concomitant medications are defined as any medication taken after the first study treatment dosing, including those who started prior to the first study treatment date and continued past that date. See Appendix 2 for handling of completely or partially missing dates for prior and concomitant medications.

Incidence of prior and concomitant medications will be tabulated by ATC level 3 and PT using the safety analysis set. A subject with the same medication taken multiple times will be counted only once for the corresponding PT. Similarly, if a subject has taken more than one medication within the same ATC level, then the subject will be counted only once for that ATC. Prior and concomitant medications will be sorted alphabetically by ATC level and within each ATC level, the PT will be presented by descending order.

11 STUDY TREATMENT EXPOSURE AND TREATMENT COMPLIANCE

A summary of exposure related to Roflumilast cream and the vehicle will be presented using the safety population by treatment group. It will include descriptive statistics on the number of days on IP, as well as the number of investigational product applications based on diary, for each treatment group. The number of days on IP, will be calculated as follows:

[(last treatment date- first treatment date) + 1].

For each subject, investigational product application compliance (%) will be calculated as follows:

$$\frac{\text{Number of investigational product applications}}{\text{Number of expected investigational product applications}} \times 100$$

Number of investigational product applications will be calculated as number of expected investigational product applications minus number of doses missed. Number of doses missed and the date that the dose was missed were collected in eCRF.

Number of expected investigational product applications will be calculated as calculated as last treatment/interruption date – first treatment date + 1. If latest treatment date ≥ latest interruption date, then the latest treatment date will be used; otherwise, latest interruption date will be used in deriving the expected number of IP applications.

Descriptive statistics for the compliance as well as the number of missed applications, subjects with < 80%, [80% - 100%], and >100% compliance will be presented by treatment. Furthermore, the incidence of subjects who missed more than 3 consecutive doses and compliant subjects will be presented by treatment.

A subject will be considered compliant with the dosing regimen if the subject meets both of the following requirements:

- applies at least 80% of the expected applications during the study drug application period
- does not miss more than 3 consecutive doses

Total weight of study medication applied (determined by weighing the study medication before and after use) will be summarized by treatment using descriptive statistics. Weight of study medication used will be documented in source documents and in eCRF. Total weight of IP used is determined by subtracting minimum of (returned, remained) tube weight from the maximum of

(dispensed, prepared) tube weight for each tube that was dispensed and summing the weights. Calculate Actual IP Used: maximum of (dispensed, prepared) tube weight minus minimum of (returned, remained) tube weight = ___ grams

If a tube is not returned at the end of the study or a tube weight is missing, actual IP used for the tube and total weight of IP applied during study will be missing.

Number of days on IP and compliance, including compliance collected at each clinic visit, will be displayed in a listing of study treatment administration by subject, for each treatment. A listing of drug accountability including the kit number, tube number, dispensed and returned weight will also be provided.

12 EFFICACY ANALYSIS

12.1 Primary Efficacy Endpoint Analysis

12.1.1 Primary Efficacy Endpoint and Estimand

The vIGA-AD is a static evaluation of qualitative overall AD severity. This global assessment scale is an ordinal scale with five severity grades (reported only in integers of 0 to 4). Each grade is defined by a distinct and clinically relevant morphologic description that minimizes inter-observer variability. vIGA-AD is evaluated for the entire body except the scalp, palms, and soles.

The primary efficacy endpoint is vIGA-AD success, defined as an vIGA-AD score of ‘clear’ (0) or ‘almost clear’ (1) plus at least a 2-grade improvement from Baseline at Week 4.

The primary estimand is described by the following attributes:

Population: Patients with Atopic Dermatitis

Endpoint: vIGA-AD success at Week 4

Intercurrent events: In the course of the 4-week randomized treatment period, subjects may be exposed to possible known or unknown intercurrent events that could possibly impact the estimates of the estimand, such as treatment discontinuation due to a specific adverse effect or perhaps a lack of effect. A composite strategy will be implemented that handles subjects who discontinue due to lack of efficacy or adverse event as missing not at random differently than all other subjects. That is, subjects who discontinue due to lack of efficacy or adverse event will be treated as non-responders for all pre-specified analysis visits (refer to Section 5.4) for which the subject’s last dose day falls within the analysis window or is prior to the start of the analysis window (refer to

Section 5.5) while the "Treatment Policy Strategy" will be adopted for handling intercurrent events in this study other than discontinuation due to lack of efficacy or adverse event.

Population-level summary: ratio of the odds of achieving vIGA-AD success after 4 weeks of using roflumilast cream 0.15%, relative to the odds of success after 4 weeks using a matching vehicle cream in the ITT population.

The supportive population-level summary: the proportion difference between Roflumilast cream 0.15% and vehicle groups will be provided for the patients who achieve vIGA-AD success at week 4 in the ITT population.

12.1.2 Hypothesis Testing

Primary hypothesis testing on the odds ratio: The null hypothesis is that the vIGA-AD success does not differ between roflumilast cream 0.15% and matching vehicle cream. The alternative hypothesis is that the vIGA-AD success does differ between roflumilast cream 0.15% and matching vehicle cream.

Null Hypothesis (H_0): $P_R\ Q_V\ /\ P_V\ Q_R = 1.0$,
Alternative Hypothesis (H_A): $P_R\ Q_V\ /\ P_V\ Q_R \neq 1.0$, where

P_R = the proportion of vIGA-AD success in roflumilast cream 0.15%
P_V = the proportion of vIGA-AD success in matching vehicle cream
$Q_R = 1 - P_R$
$Q_V = 1 - P_V$.

12.1.3 Primary Endpoint Analysis

For the primary analysis, missing vIGA-AD scores will be imputed using multiple imputation as described in Section 6.2.1. These imputations will result in a minimum of 25 to a maximum of 250 complete analysis datasets, depending on the number of imputed monotone datasets that are required.

Percentages of subjects having a vIGA-AD Success (refer to Section 12.1.1) will be presented by visit and treatment group based on multiply imputed data in the ITT population along with a 95% Wilson CI. The common MH odds ratio and common MH proportion difference, adjusted for the randomization factors (i.e., randomized vIGA-AD score and pooled study site) will also be provided along with their common associated 95% CIs. Additionally, count and percentage of subjects having a vIGA-AD success, count and percentage of subjects in each category of the vIGA-AD scale, and descriptive statistics for the vIGA-AD scores, change in vIGA-AD score from

baseline and percent change in vIGA-AD score from baseline will be presented by visit and treatment group based on observed data in the ITT population.

The primary endpoint (vIGA-AD Success at Week 4) will then be analyzed using multiple imputation CMH test stratified by the randomization factors. To do so, a CMH analysis will be performed separately for each of the complete multiply imputed analysis data sets, and results will be combined into one multiple imputation inference using the methodology described in Section 6.2.1. Statistical significance will be concluded at the 5% significance level (2-sided). Should the odds ratio be not estimable for at least one multiply imputed dataset, the conclusion of the study will be based on the p-value obtained from a MH test, stratified by the randomization factors, for the common MH proportion difference at Week 4.

The following sensitivity analyses to the primary analysis of the primary endpoint will be performed:

- Multiply-imputed data (refer to Section 6.2.1) on mITT population (refer to Section 4.2) and
- Tipping Point analysis (refer to Section 6.2.3) on ITT population.
- Non-responder imputation (refer to Section 6.2.2) on ITT population.
- Observed data on ITT population (refer to Section 4.1).
- Observed data on PP population (refer to Section 4.3).

For the last three sensitivity analyses, count of subjects having vIGA-AD success will also be presented by visit and treatment group in addition to the percentage of subjects having vIGA-AD success.

To assess the impact of the pooling of the study sites on the primary analysis of the primary endpoint, the following analyses will be performed:

- The primary analysis of the primary efficacy endpoint will be repeated but with a different site pooling strategy (refer to Section 6.4).
- To assess the impact of site on the primary analysis endpoint, the proportion of subjects achieving vIGA-AD success and 95% CI within each site will be tabulated. No p-values will be provided. Forest plots of the proportions (and associated 95% CI) for each site and treatment will also be provided.

- An additional analysis to examine the impact of study site will examine the changes in p-values that occur after removal of subjects from a site. To do so, the primary analysis of the primary efficacy endpoint will be repeated but removing a different pooled study site for each iteration. Forest plots of the odds ratio (and associated 95% CI) for each pooled site removed will also be provided.

12.2 Secondary Endpoints Analysis

Secondary efficacy endpoints are based on the vIGA-AD scale, WI-NRS scale or EASI questionnaire. The list of secondary efficacy endpoints can be found in Section 2 and their derivation in Section 5.5.

- For more details about the vIGA-AD scale refers to Section 12.1.1.
- The WI-NRS scale is a single item scale assessing the subject-reported worst itch severity during the previous 24-hour period. The scale is from 0 to 10 ('no itch' to 'worst itch imaginable').
- For the EASI questionnaire, four anatomic sites (head, upper extremities, trunk, and lower extremities) are assessed for erythema, induration/infiltration (papules), excoriation, and lichenification as seen on the day of the examination. The severity of each sign is assessed using a 4-point scale (half steps are allowed e.g., 0.5): 0 (none), 1 (mild), 2 (moderate), and 3 (severe). The area affected by AD within a given anatomic site is estimated as a percentage of the total area of that anatomic site and assigned a numerical value according to the degree of AD involvement as follows: 0 (no involvement), 1 (1-9%), 2 (10-29%), 3 (30-49%), 4 (50-69%), 5 (70-89%), and 6 (90-100%).

For each secondary efficacy endpoint, missing data will be imputed and data will be summarized as for the primary efficacy endpoint (refer to Section 12.1.3).

Upon successful demonstration of statistical significance for the primary efficacy endpoint, the secondary efficacy endpoint of vIGA-AD success among subjects with vIGA-AD score of 'Moderate' at randomization will be analyzed as described for the primary efficacy endpoint (refer to Section 12.1.3) but based on the vIGA-AD Moderate ITT population (refer to Section 4.4) and stratifying by pooled study site only. That is, as per the vIGA-AD Moderate ITT population definition all subjects included in this analysis will have a randomized vIGA-AD score of Moderate (3) and so, the CMH test cannot be stratified by randomized vIGA-AD score.

Upon successful demonstration of statistical significance for the primary and above secondary efficacy endpoints, the remaining secondary efficacy endpoints will be grouped into secondary

endpoint family 1, comprised of the 4-point reduction on the average weekly WI-NRS at Week 4, Week 2 and Week 1, and secondary endpoint family 2, comprised of the EASI-75 at Week 4, vIGA-AD of 'clear' or 'almost clear' at Week 4, vIGA-AD of success at week 2 and week 1, vIGA-AD of 'clear' or 'almost clear' at Week 2 and Week 1. The analysis of secondary endpoint family 1 will be performed as described for the primary efficacy endpoint (refer to Section 12.1.3) but based on the WI-NRS population (refer to Section 4.6) and an alpha level of 0.03 will be used to test these endpoints sequentially as described in Section 6.5. Similarly, the analysis of secondary endpoint family 2 will be performed as described for the primary efficacy endpoint (refer to Section 6.2.1) based on the ITT population (refer to Section 4.1) but an alpha level of 0.02 will be used to test these endpoints sequentially as described in Section 6.5.

Secondary Endpoint Family 1 (α=0.03, hierarchical testing)

- In subjects ≥12 years old with baseline WI-NRS ≥ 4, achievement of at least a 4-point reduction in the average weekly WI-NRS at Week 4
- In subjects ≥12 years old with baseline WI-NRS ≥ 4, achievement of at least a 4-point reduction in the average weekly WI-NRS at Week 2
- In subjects ≥12 years old with baseline WI-NRS ≥ 4, achievement of at least a 4-point reduction in the average weekly WI-NRS at Week 1

Secondary Endpoint Family 2 (α = 0.02, hierarchical testing)

- Achievement of at least a 75% reduction in the Eczema Area and Severity Index at Week 4 (EASI-75)
- vIGA-AD of 'clear' or 'almost clear' at Week 4
- vIGA-AD Success at Week 2
- vIGA-AD Success at Week 1
- vIGA-AD of 'clear' or 'almost clear' at Week 2
- vIGA-AD of 'clear' or 'almost clear' at Week 1

For the secondary efficacy endpoint of vIGA-AD Success at Week 4 based on the vIGA-AD moderate ITT population, the following sensitivity analyses to the primary analysis of this efficacy endpoint will be performed if and only if the hierarchical testing procedure (refer to Section 6.5) allows it:

- Tipping Point analysis (refer to Section 6.2.3) on vIGA-AD moderate ITT population.

- Non-responder imputation (refer to Section 6.2.2) on vIGA-AD moderate ITT population.
- Observed data on vIGA-AD moderate ITT population.
- Observed data on vIGA-AD moderate PP population (refer to Section 4.5).

For the last three sensitivity analyses, count of subjects having vIGA-AD success will also be presented at Week 4 by treatment group in addition to the percentage of subjects having vIGA-AD success.

For all other secondary efficacy endpoints, the following sensitivity analyses to the primary analysis of these secondary endpoints will be performed if and only if the hierarchical testing procedure (refer to Section 6.5) allows it:

- Non-responder imputation on WI-NRS population for the WI-NRS secondary efficacy endpoints and ITT population for all other secondary efficacy endpoints.
- Observed data on WI-NRS population for the WI-NRS secondary efficacy endpoints and ITT population for all other secondary efficacy endpoints.

The daily WI-NRS score, average weekly WI-NRS, and WI-NRS success flag will be listed.

The EASI total score and EASI-75 flag at each visit will be listed.

12.3 Exploratory Efficacy Endpoints

Exploratory efficacy endpoints are based on the vIGA-AD scale, WI-NRS scale, EASI questionnaire, %BSA affected by AD, CDLQI/DLQI questionnaires, DFI questionnaire, SCORAD tool or POEM tool. The list of exploratory efficacy endpoints can be found in Section 2 and their derivation in Section 5.5.

- For more details about the vIGA-AD scale, refer to Sections 12.1.1.
- For more details about the WI-NRS scale, refer to Section 12.2.
- For more details about the EASI questionnaire, refer to Section 12.2.
- The % of BSA affected by AD will be determined by the subject's hand method, where the subject's hand (including fingers) surface area is assumed to equal 1% of BSA (excluding the scalp, palms, and soles).
- The CDLQI/DLQI is a self-administered validated questionnaire designed to measure the health-related quality of life of children/adult subjects suffering from a skin disease. It

consists of 10 questions concerning subjects' perception of the impact of skin disease on different aspects of their health-related quality of life over the last week. Questions 1 to 6 and 8 to 10 are rated from 0 (Not at all) to 3 (Very much). For children, if last week was a school time, question 7 is rated from 0 (Not at all) to (Prevented school) but if the last week was a holiday time, question 7 is rated from 0 (Not at all) to 3 (Very much). For adults, question 7 is rated from 0 (Not at all/Not relevant/No) to 3 (Yes).

- The DFI questionnaire measures how much having a child with AD affects the quality of life of other (adult) members of the family over the last week. It is designed to be completed by caregivers of subjects ≤ 17 years of age and consists of 10 questions rated from 0 (Not at all) to 3 (Very much).

- The SCORAD is a clinical tool to assess the severity (i.e., extent, intensity) of AD as objectively as possible. First, the overall %BSA affected by AD is evaluated (from 0% to 100%, where a subject's palm represents 1% of his/her total BSA). Secondly, the AD severity is evaluated based on 6 items (erythema, edema/papulation, oozing/crusts, excoriation, lichenification, and dryness) graded using a 4-point scale (half steps are not allowed): 0 (absence), 1 (mild), 2 (moderate), and 3 (severe). Lastly, 2 subjective items (loss of sleep and intensity of pruritus) are evaluated by having the subject indicates on a 10.0 cm visual analog scale (VAS) the point corresponding to the average value over the last 3 days (0 cm = none to 10 cm= maximum).

- The POEM is a tool used for monitoring atopic eczema severity. If focus on the illness as experience by the subject. It consists of a 5-point scale measuring the frequency of each of 7 AD symptoms (i.e., dryness, itching, flaking, cracking, sleep loss, bleeding, and weeping) over the past week scored from 0 (no days), 1 (1 to 2 days), 2 (3 to 4 days), 3 (5 to 6 days), and 4 (every day).

For the continuous exploratory efficacy endpoints of change and percent from baseline in average weekly WI-NRS, EASI total score, % BSA, DLQI/CDLQI score, DFI score, SCORAD score, and POEM score at Week 1, Week 2, and Week 4, descriptive statistics for the score, change from baseline and percent change from baseline will be presented by visit and treatment group based on observed data in the ITT population. Descriptive statistics will also be presented similarly for the average weekly WI-NRS at Week 3. For the continuous efficacy exploratory endpoint of daily WI-NRS, descriptive statistics will be presented similarly but on a daily basis instead of a weekly basis. Additionally, a plot of the mean (and standard error) daily WI-NRS scores over time for each treatment group will also be provided based on observed data. A similar plot will be provided for the percent change from baseline in daily WI-NRS.

For each categorical exploratory efficacy endpoint, count and percentages of subjects meeting the criteria for an exploratory efficacy endpoint will be presented by visit and treatment group based on observed data in the ITT population along with a 95% Wilson CI. Additionally, count and percentage of subjects meeting the criteria for vIGA-AD of Clear (0), EASI-50, EASI-90, and EASI-100 at Week 1 and Week 2 will also be presented by visit and treatment group based on observed data in the ITT population along with a 95% Wilson CI.

Continuous exploratory efficacy endpoints will be analyzed at Week 1, Week 2, and Week 4 and for daily WI-NRS assessments using an analysis of covariance (ANCOVA) with the factors of treatment, two stratification variables (pooled study site and randomized vIGA-AD score), and baseline of the variable under analysis as covariate. Statistical comparisons between the treatment groups will be obtained using contrasts. The Least Square (LS) mean and its standard error, difference in LS means between treatment group (i.e., active - vehicle), its standard error and associated 95% confidence interval, and p-value for difference from vehicle will be presented at each visit. These analyses will be performed based on observed data in the ITT population.

Categorical exploratory efficacy endpoints at Weeks 1, 2, and 4 will be analyzed using a CMH test adjusted for the two stratification variables (pooled study site and randomized vIGA-AD score). Common MH odds ratio, common MH proportion difference and their common associated 95% CI, adjusted for the randomization factors, will be provided. The p-value will be from a CMH test for the common MH odds ratio, unless the common MH odds ratio is not estimable. In such circumstances, the p-value will be from a MH test for the common MH proportion difference.

The analysis of these endpoints will be performed on the observed data with no imputation. The p-values will be nominal as no formal inferential testing will be done on exploratory efficacy endpoints.

12.4 Subgroup Analysis

With the following exception, analyses of the primary and secondary efficacy endpoints (refer to Sections 12.1 and 12.2, respectively) will be repeated by subgroups (refer to Section 6.6) based on the multiple imputation data using mITT, ITT or WI-NRS population, as applicable (refer to Section 12.2). The exception is that the subgroup analyses by randomized vIGA-AD score and by actual baseline vIGA-AD score will not be performed for the secondary efficacy endpoint of vIGA-AD Success based on the vIGA-AD Moderate ITT population.

The following alternatives could be implemented if the odds ratio and/or proportion difference are not estimable:

1. An unstratified model can be used if odds ratios are not estimable due to over-stratification, i.e., too few events for the number of strata.
2. If the between-imputation variance of odds ratios and/or proportion difference is 0, i.e., the estimates are equal across all imputations, the estimates from any imputation will be reported.

Should a specific subgroup have less than 10 subjects across both treatment groups, no statistical inference will be performed. Forest plots of subgroup analysis for primary efficacy endpoint and secondary efficacy endpoints will also be provided.

No subgroup analyses will be presented for the exploratory efficacy endpoint (refer to Section 2).

12.5 Summary of Primary and Secondary Efficacy Analysis

Table 5 provides a summary of the primary and sensitivity analyses that will be provided for primary and secondary efficacy endpoints.

Table 6 Summary of Primary and Secondary Efficacy Analyses

Efficacy Endpoint	**Primary Analysis**	**Sensitivity Analysis**
Primary (refer to Section 12.1)		
vIGA-AD Success (a score of '0' or '1' plus at least a 2-grade improvement) at week 4	ITT, multiple imputation (CMH)	#1 mITT, multiple imputation (CMH) #2 ITT, Tipping point (CMH)) #3 ITT, non-responder imputation (CMH) #4 ITT, observed data (CMH) #5 PP, observed data (CMH) #6 ITT, multiple imputation (CMH) and a different site pooling strategy (CMH) #7 ITT, multiple imputation (CMH) by pooled study site #8 ITT, multiple imputation (CMH) and removing one pooled study site at the time
Secondary (refer to Section 12.2)		

vIGA-AD Success (a score of '0' or '1' plus at least a 2-grade improvement) at week 4	vIGA-AD Moderate ITT, multiple imputation (CMH)	#1 vIGA-AD Moderate ITT, Tipping point (CMH) #2 vIGA-AD Moderate ITT, non-responder imputation (CMH) #3 vIGA-AD Moderate ITT, observed data (CMH) #4 vIGA-AD Moderate PP, observed data (CMH)
Family 1 (α=0.03): Average weekly WI-NRS Success (achievement of at least a 4-point reduction) at weeks 1,2, and 4	WI-NRS population, multiple imputation (CMH)	#1 WI-NRS, non-responder imputation (CMH) #2 WI-NRS, observed data (CMH)
Family 2 (α=0.02 or α=0.05): EASI-75 at week 4	ITT, multiple imputation (CMH)	#1 ITT, non-responder imputation (CMH) #2 ITT, observed data (CMH)
Family 2 (α=0.02 or α=0.05): vIGA-AD Success (a score of '0' or '1' plus at least a 2-grade improvement) at weeks 1 and 2	ITT, multiple imputation (CMH)	#1 ITT, non-responder imputation (CMH) #2 ITT, observed data (CMH)
Family 2 (α=0.02 or α=0.05): vIGA-AD score of '0' or '1' at weeks 1, 2 and 4	ITT, multiple imputation (CMH)	#1 ITT, non-responder imputation (CMH) #2 ITT, observed data (CMH)

13 SAFETY ANALYSIS

Safety analyses will be conducted using the safety population. Subjects will be analyzed based on the treatment received and the stratum they belong to.

No formal inferential statistics will be performed on safety assessments.

Also, treatment-emergent adverse events by SOC and PT table and summary of weight change will be presented by treatment groups for age 6 to 17 years and ≥ 18 years separately.

13.1 Adverse Events

Adverse events (AEs) will be coded according to the latest available version of Medical Dictionary for Regulatory Activities (MedDRA) (Version 25.0).

Treatment emergent adverse events (TEAEs) are defined as any AEs with onset on or after the first study drug application. See Appendix 2 for handling of completely or partially missing dates for AEs. In the case where it is not possible to define an AE as treatment emergent or not, the AE will be classified as treatment emergent. All reported TEAEs will be summarized by treatment group.

Overall summary will be presented, which will include the total number of events, and the number and percentage of subjects who experienced TEAE, TEAE by the strongest relationship, TEAE by the maximum severity, treatment-related TEAE by maximum severity, treatment-emergent serious AE (TESAE), treatment-emergent Non-SAE, TEAE leading to study treatment discontinuation, TEAE leading to study discontinuation, TEAE on an application site, and TEAE leading to death.

The number and percentage of subjects who experience TEAE will be summarized by SOC and PT within SOC. In addition, similar table will be presented by treatment groups for age 6 to 17 years and ≥ 18 years separately. Unless otherwise specified, a subject experiencing the same TEAE multiple times will be counted only once for the corresponding PT. Similarly, if a subject experiences multiple TEAEs within the same SOC, the subject will be counted only once for that SOC. TEAEs will be sorted alphabetically by SOC and within each SOC the PT will be presented by descending frequency in the safety analysis set. A treatment-related TEAE is defined as any TEAE that is assessed by the Investigator as likely, probably, or possibly related to study treatment. TEAE that is assessed as unrelated or unlikely will be defined as not treatment-related. If a subject experiences more than one TEAE within different relationship categories within the same SOC/PT, only the worst case (the strongest relationship) will be reported. TEAE with an unknown relationship will be considered as treatment-related.

The number and percentage of subjects who experience TEAE will be summarized by SOC, PT and the maximum severity (mild/moderate/severe/life threatening/death related to AE). If a subject experiences more than one TEAE within different severity categories within the same SOC/PT, only the worst case (the maximum severity) will be reported. TEAE with an unknown severity will be considered as severe.

The number and percentage of subjects who had TESAE, treatment-emergent Non-SAE, will be summarized by SOC and PT within SOC.

Frequency and percentage of subjects who experience TEAE on an application site will be summarized by SOC and PT.

A table and plot of most frequent TEAE (‘≥1%) by PT will be provided by treatment arms (overall TEAE, overall TESAE will be included in the same plot).

All the AEs will be listed. Any TEAE leading to death will also be included in the AE listing (if there is any). The TEAE related to application site will be flagged in the AE listing.

13.2 Clinical Laboratory

Descriptive statistics for the observed values in chemistry, hematology, and quantitative urinalysis, change from baseline and percent change from baseline values will be presented by treatment group at each scheduled visit. For qualitative urinalysis data, the number and percentage of the subjects for each level of result by treatment at each scheduled visit will be provided.

Shift tables from baseline to each postbaseline assessments describing shifts to out-of-normal range will be provided for chemistry, hematology, and qualitative urinalysis. Only subjects with a baseline result and a result at the specified visit for the parameter will be considered.

Listings of abnormal laboratory will be provided for each parameter where a subject had at least one abnormal result.

Laboratory data will be presented in SI units.

13.3 Vital Signs

Descriptive statistics will be presented for vital signs (systolic blood pressure, diastolic blood pressure, heart rate, body temperature, and weight). Observed values, change from baseline and percent change from baseline values will be presented by treatment group at each scheduled visit.

The number and percentage of subjects with gain or lose >5% from baseline in body weight over the course of the study will be summarized for overall and separately for children and adolescents (6 to 17 years) and adults (≥ 18 years). BMI will be summarized using descriptive statistics (n, mean, standard deviation, median, Q1, Q3, minimum, and maximum) on observed values, change from baseline and percent change from baseline values separately for children and adolescents (6 to 17 years) and adults (≥ 18 years). The BMI percentile rather than BMI kg/m^2 will be summarized for children and adolescents (6 to 17 years).

A listing of all vital sign assessments including weight, BMI, and BMI percentile for children and adolescents will be provided.

13.4 Local Tolerability Assessments

The investigator's assessment of the application site reaction will be summarized by visit using both categorical methods (number and percentage of subject with each score) as well as continuous methods (e.g., mean, standard deviation, etc.)

Local tolerability (burning/stinging sensation) assessed by the subject will be summarized using number and percentage similarly.

13.5 Patient Health Questionnaire Depression Scale (PHQ-8) and Patient Health Questionnaire Depression Scale (Modified PHQ-A)

The Modified PHQ-A Assessment will be performed in adolescent subjects (12-17 years old, inclusive; question 9 has been removed since that is better evaluated by use of the C-SSRS tool).

The PHQ-8/Modified PHQ-A score is the sum of the responses for the questions, each question ranging from 0 (Not at all) to 3 – (Nearly every day). Five severity categories of depression are defined as follows:

- None – Minimal depression (0 to 4)
- Mild depression (5 to 9)
- Moderate depression (10 to 14)
- Moderately severe depression (15 to 19)
- Severe depression (20 to 24)

The score will be set to missing in case of at least one missing value. The number and percentage of subjects in each category will be summarized by treatment and visit. A summary of the shifts in depression category from baseline to each study visit will also be provided.

13.6 Children's Depression Inventory 2 (CDI-2)

The CDI-2 Assessment will be performed in children subjects (6-11 years old, inclusive).

The observed values and changes from baseline will be calculated for CDI-2 total score and the 2 scales, i.e., emotional problems and functional problems, and will be summarized descriptively by treatment group and visit.

The CDI-2 total score will be categorized as follows:

1. Normal: Male Score < 22, Female Score < 21
2. Elevated: $22 \leq$ Male Score < 32, $21 \leq$ Female Score < 32
3. Very Elevated: Male or Female Score ≥ 32.

The proportion of subjects that meeting the criteria will be summarized by treatment group and visit.

A summary of the shifts in CDI-2 category from baseline to each study visit will also be provided.

13.7 Columbia-Suicide Severity Rating Scale (C-SSRS)

The Columbia-Suicide Severity Rating Scale (C-SSRS) is a questionnaire used (for adolescents and adults 12 years old and older) for suicide assessment developed by multiple institutions, including Columbia University. The C-SSRS prospectively assesses Suicidal Ideation and Suicidal Behavior. At the Screening study visit, "Baseline/Screening" version of the C-SSRS will be used. This version assesses Suicidal Ideation and Suicidal Behavior during the subject's lifetime and during the past 6 months. For the Screening visit, "lifetime" experience of the subject with Suicidal Ideation and Suicidal Behavior will be summarized. On all subsequent visits, the "Since Last Visit" version will be used (Baseline/Day 1, Week 1/Day 8, Week 2/Day 15 or Week 4/Day 29).

Tables will include results from the Suicidal Ideation and Suicidal Behavior sections of the C-SSRS. Frequencies and percentages of subjects with a response of "Yes" at any point on the Suicidal Ideation and Suicidal Behavior items will be summarized by study visit and treatment group.

13.8 Physical Examination

The number and percentage of subjects with normal and abnormal findings in the physical examination will be presented by body system and treatment group at each study visit.

14 PHARMACOKINETICS ANALYSIS

Concentration data will be summarized by visit by age group (<18 vs. >=18) and overall, for active treatment group using descriptive statistics, reporting n, mean, standard deviation, median, Q1, Q3, minimum, and maximum, and geometric statistics including geometric mean and coefficient of variation. For computation of mean plasma concentrations, data that are below the limit of quantification (BLQ) will be set to 0.0001. The PK population will be used for these analyses.

PK data will be presented in the listing.

15 REFERENCES

1. Anne Lott & Jerome P. Reiter (2020) Wilson Confidence Intervals for Binomial Proportions with Multiple Imputation for Missing Data, The American Statistician, 74:2, 109-115, DOI: 10.1080/00031305.2018.1473796

2. Ratitch, B., Lipkovich, I., & O'Kelly, M. (2013). *Combining Analysis Results from Multiply Imputed Categorical Data*. PharmaSUG. https://www.pharmasug.org/proceedings/2013/SP/PharmaSUG-2013-SP03.pdf

16 APPENDICES

Appendix 1

Output Conventions

TLF will be generated using SAS® and will be displayed on letter size paper with landscape orientation, 1-inch margins and 9 pt. Courier New font.

The header section will comprise the sponsor's name, the protocol number, the delivery description, the data cut-off date (if applicable), the TLF number, the TLF title, the analysis set, and the page number (Page X of Y). The footer section will include the TLF footnotes, the CRO's name, the date and time of the execution of the program, and the name of the program.

P-values equal to or above 0.0001 will be reported to 4 decimal places; p-values less than 0.0001 will be reported as "<0.0001"; p-values greater than 0.9999 will be reported as ">0.9999".

The mean, median, geometric mean will be displayed to one more decimal place than the original value; Q1, Q3, minimum and maximum will keep the same number of decimal places as the original value; standard deviation, standard error, CV and CI will be displayed to two more decimal places than the original value. If derived parameters are to be summarized, the number of decimals of the derived values is to be chosen on a case-by-case basis, but the rule above applies.

For categorical summary tables, percentages will be reported to one decimal place. Percentages between 0 and 0.1 (both exclusive) will be displayed as "<0.1". The denominator for each percentage will be the number of subjects within the population per treatment group unless otherwise specified.

Listings will be ordered by treatment group, subject number, date and visit (where applicable). Imputed dates will not be presented in the listings.

Dates & Times Format

Date and time (if available) will be presented in the format yyyy-mm-dd/hh:mm.

Presentation of Treatment Groups

When applicable, study treatments will be represented as follows in the different outputs:

Study Treatment Full Name	Study Treatment Output Name
Roflumilast cream 0.15%	Roflumilast Cream 0.15%
Vehicle cream	Vehicle

Appendix 2

Algorithm for Imputation of Start/End Date and Time of Adverse Events and Prior/Concomitant Medications

Event Start Date Imputation

- Imputation of event end date should be done before imputation of event start date.
- Completely missing: Impute to the first study treatment date.
- Missing day and month: Impute to January 1st, unless year is the same as year of first study treatment dose date then impute to the first study treatment date.
- Missing day: Impute to the 1st day of the month, unless month and year are the same as month and year of first study treatment dose date then impute to the first study treatment date.
- If imputed event start date is after event end date (imputed or not), set the event start date to the imputed event end date.

Event Start Time Imputation (for Adverse Events only)

Imputation of event end time should be done before imputation of event start date.

- If the event date is not the same as the first dose date or time part of the first dose date is missing, impute to 00:00.
- If the event date is the same as the first dose date and event occurred prior to study drug application (as flagged in CRF), impute to 00:00.
- If the event date is the same as the first dose date and event did not occur prior to study drug application (as flagged in CRF), impute to time part of first dose date.
- If the event start date is equal to event end date and imputed event start time is after event end time (imputed or not), set the event start time to the imputed event end time.

Event End Date Imputation

- Completely missing (and not flagged as "ongoing"): Impute to the last contact date.
- Missing day and month: Impute to December 31st, unless year is the same as last contact date then impute to the last contact date.
- Missing day: Impute to the last day of the month, unless year and month are the same as year and month of last contact date then impute to the last contact date.

Event End Time Imputation (for Adverse Events only)

Impute to 23:59.