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 0.05% Administered QD in Subjects with Atopic Dermatitis**
Protocol Number and Version:	**ARQ-151-315, Original dated December 13, 2020** **Amendment 1 dated July 16, 2021** **Amendment 2 dated April 10, 2023**
Product:	**ARQ-151 Cream 0.05%**
Sponsor:	**Arcutis Biotherapeutics, Inc.** **3027 Townsgate Road, Suite 300** **Westlake Village, CA 91361**
Date:	**12 April 2023**
Version:	**1.0**
Prepared by:	**Innovaderm Research Inc.** **3530 boul. Saint-Laurent, suite 300,** **Montreal, Quebec, Canada, H2X 2V1**

STATISTICAL ANALYSIS PLAN REVISION SUMMARY			
Version	**Version Date**	**Author**	**Summary of Changes**
Original V1.0	12-APR-2023		Initial version

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)
	Principal Biostatistician Innovaderm Research		

APPROVALS:

Name	Title	Signature	Date (DD-MMM-YYYY)
	Associate Director, Biostatistics Innovaderm Research		
	Vice President, Biometrics Arcutis Biotherapeutics, Inc.		
	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
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
CSR	Clinical Study Report
DFI	Dermatitis Family Impact
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
IDQoL	Infant Dermatology Quality of Life Index
IP	Investigational Product
ITT	Intent to Treat
IWRS	Interactive Web Response System
LoE	Lack of efficacy
MCMC	Markov-Chain Monte-Carlo
MedDRA	Medical Dictionary for Regulatory Activities
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-151-315. The analyses described in the SAP are based upon the protocol Amendment 2 dated 10 April 2023. 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.

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 ARQ-151 cream 0.05% vs vehicle administered once daily (QD) x 4 weeks to subjects aged 2-5 years with atopic dermatitis (AD).	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) 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 • 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 of 'clear' or 'almost clear' at Week 2 • vIGA-AD Success at Week 1

OBJECTIVES	ENDPOINTS
	• vIGA-AD of 'clear' or 'almost clear' at Week 1
	Exploratory efficacy endpoints:
	• Exploratory 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, Children's Dermatology Life Quality Index (CDLQI)/Infant's Dermatitis Quality of Life Index (IDQoL), the Dermatitis Family Impact (DFI), Scoring Atopic Dermatitis (SCORAD), and Patient-oriented Eczema Measure (POEM) at Week 1, Week 2 and Week 4. • Exploratory categorical efficacy endpoints include EASI-75 at Week 1 and Week 2; EASI-50, EASI-90, EASI-100, achievement of at least a 4-point reduction on the WI-NRS, and vIGA-AD of 'clear' at Week 1, Week 2 and Week 4.
Safety	
To assess the safety of ARQ-151 cream 0.05% vs vehicle administered QD x 4 weeks to subjects aged 2-5 years with AD.	• Incidence and severity of adverse events (AEs) • Changes and percent change in vital signs • The subject incidence of >=5% weight loss or gain on study • Local tolerability assessments

3 STUDY DESIGN

3.1 Overall Design

This is a Phase 3, parallel group, double blind, vehicle-controlled study in which ARQ-151 cream 0.05% or vehicle is applied QD x 4 weeks to subjects with mild to moderate atopic dermatitis.

At entry, subjects will have ≥3% BSA involvement (excluding the scalp, palms, soles) and mild or moderate atopic dermatitis based on vIGA-AD assessment.

Upon determination of eligibility, subjects will be randomized 2:1 to either ARQ-151 cream 0.05% or matching vehicle cream. The randomization will be stratified by vIGA-AD score at baseline ('Mild' vs. 'Moderate'), and by study site.

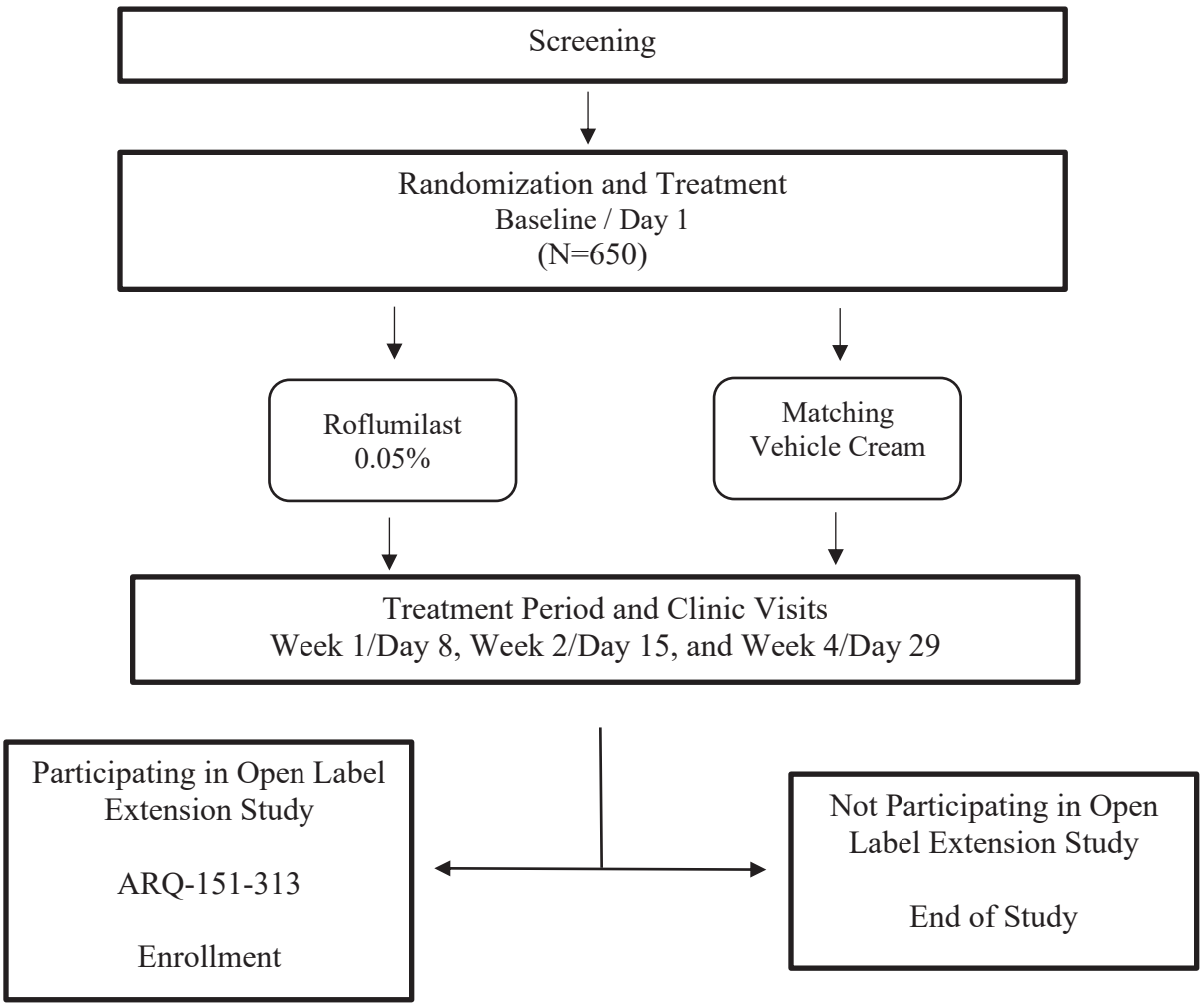
Caregivers will apply ARQ-151 cream 0.05% or vehicle cream QD to all AD affected areas and any newly appearing AD lesions that arise during the study, except on the scalp. Caregivers should maintain treatment of these areas with study drug for the duration of the study regardless of whether treatable areas of AD clear prior to Week 4/Day 29.

Subject participation involves a minimum of 5 clinic visits including Screening, Baseline/Day 1, Week 1/Day 8, Week 2/Day 15, and Week 4/Day 29. The interval between the Screening and Baseline/Day 1 visits could be up to 30 days, therefore the anticipated maximum duration of subject participation is approximately 8 weeks.

At the Week 4/Day 29 visit, subjects may be eligible to enroll in a 12-month, open label extension study (ARQ-151-313) evaluating ARQ-151 cream 0.15% QD and 0.05% QD. All subjects will have the option to enroll into the ARQ-151-313.

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 0.05% Administered QD in Subjects with Atopic Dermatitis

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

- Roflumilast cream 0.05% 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

3.2 Schedule of Events

Table 1 Schedule of Visits and Assessments

Study Procedure	Screen	Baseline Day 1	Week 1 Day 8	Week 2 Day 15	Week 4 Day 29/ET
Visit	**1**	**2**	**3**	**4**	**5**
Visit Window	**-30 days**	**N/A**	**± 3 days**	**± 3 days**	**± 3 days**
Parents/Caregivers informed consent	X				
Demographics	X				
Medical and surgical history	X				
Physical examination[a]	X	X			X
I/E criteria	X	X			
Hematology and Serum Chemistries[b]	X				
Vital signs, weight, height[c]	X	X	X	X	X
vIGA-AD[d], EASI[d], BSA[d], 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 Assessments[g]		X	X	X	X
CDLQI[h], IDQoL[h], DFI[h]	X	X	X	X	X
Medical Photography[i]		X	X		X
Drug application and Parents/Caregivers training in clinic[j]		X	X	X	
Dispense/Redispense investigational product kit[k]		X	X	X	X[l]
Weigh investigational product tubes[l]		X	X	X	X
Dispense/review diary	X	X	X	X	X
Compliance determination[m]			X	X	X
Adverse event assessment[n]	X	X	X	X	X
Concomitant medications	X	X	X	X	X
Study Exit[o]					X

Footnotes from above table:

a. Limited physical examination: skin (including assessment of Fitzpatrick skin type at Screening only), lungs, and heart only.
b. To be collected at Screening. In addition, laboratory safety tests may be performed at various unscheduled time points, if deemed necessary by the Investigator.
c. Height will be collected at Screening and Week 4/Day 29. 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). Measure the weight in triplicate and report the average weight in EDC. A 5% weight loss (whether or not intentional or other explained) should be reported to the medical monitor in a timely manner and before the next clinic visit.
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. Parents/guardians/caregivers of children will report pruritus at home on a daily basis starting 7 days prior to the Baseline/Day 1 visit, and then every day thereafter using a diary. WI-NRS score will be determined by 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 each day. Caregiver/parents will be given instructions on how to complete this questionnaire.
f. POEM will be completed for all subjects by proxy completion.
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) and 10-15 minutes post-drug application for the Subject's '0-3' burning/stinging assessment. Parents/guardians will complete assessment for subjects. Subject's burning stinging assessment at Week 4/Day 29 will be provided by a recall assessment of burning/stinging experienced post drug application on the previous day (Day 28). **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.** If a subject is entering ARQ-151-313 Open Label Extension study, then the Subject's '0-3' burning/stinging assessment will be performed as described in the ARQ-151-313 protocol for Baseline/Day 1.
h. The CDLQI will be completed by parents/caregivers for subjects ≥4 years of age. The IDQoL will be completed by parents/caregivers for subjects <4 years. The Dermatitis Family Impact Questionnaire (DFI) will be completed by parents/caregivers for all subjects.
i. 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.
j. Subjects to apply assigned IP during clinic visits, except for the Day 29/ET visit. If a subject is entering ARQ-151-313 Open Label Extension study, then subjects will apply IP as described in the ARQ-151-313 protocol for Baseline/Day 1.
k. It is expected that kits will be dispensed based on %BSA affected. See IP Handling Manual for details. 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 using Unscheduled Visits or home delivery of IP, 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 (contact sponsor representative for guidance).
l. Every tube should be weighed and recorded when dispensed and returned. See IP Handling Manual for details.

[m.] Compliance determination is described in the IP Handling Manual.

[n.] 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 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 value(s) will be evaluated by the Principal Investigator (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-151-315) will only be followed until they exit from this study). Subjects who enroll into the open label extension study (ARQ-151-313) must complete the ARQ-151-315 Week 4/Day 29 visit as this visit (Week 4/Day 29) is the Day 1 visit for ARQ-151-313.

3.3 Treatment

Roflumilast cream 0.05% or vehicle cream will be administered QD for 28 days (+/- 3 days).

- Roflumilast cream 0.05%
- Vehicle cream

3.4 Randomization, Replacement, and Unblinding Procedures

Randomization will take place at the Baseline/Day 1 visit after the Investigator confirms that the subject meets all eligibility criteria. Subjects will be randomly assigned to apply ARQ-151 cream 0.05 % QD, or matching vehicle QD. Assignment of drug or vehicle will be made at a 2:1 ratio (drug:vehicle) and stratified by Baseline/Day 1 vIGA-AD score (‘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 Interactive Web 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.

Subjects that withdraw or are discontinued from the study will not be replaced.

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

Not applicable

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 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 randomized stratum. Actual and randomized treatment will only differ if the subject received the wrong treatment throughout their participation in the study.

4.3 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.

4.4 WI-NRS Population

WI-NRS population will be a subset of the ITT population who:

1. Parent/caregiver 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 $\geq$ 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 baseline score will be treated as missing.

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

4.5 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.

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 Appendix 1).

5.1 Sample Size

Approximately 650 subjects are planned to be randomized in this study. To test the secondary endpoint of IGA 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 study site.

This study is designed to differentiate between a vehicle cream treatment that has an IGA success rate of 22% or less and ARQ-151 cream treatment with an IGA success rate of 37% or more. The hypothesis is that the IGA success rate among subjects treated with ARQ-151 cream 0.05% is significantly greater than the IGA success rate among subjects treated with vehicle cream.

This sample size provides more than 90% power to detect 15% difference between treatment groups on vIGA-AD success at Week 4 at α=0.025 using a 2-sided stratified Cochran-Mantel-Haenszel (CMH) test. The results from a recent phase 2 study (ARQ-151-212) of ARQ-151 cream 0.05% compared to vehicle treatment were used to estimate the treatment difference. The vIGA-AD success rate at Week 4 was approximately 38% in the ARQ-151 cream 0.05% group compared to 22% in the vehicle. Similarly, approximately 37% of subjects demonstrated vIGA-AD success at Week 4 in the ARQ-151 cream 0.15% group. This sample size also provides more than 80% power to detect an overall 15% difference between treatment groups on IGA success at Week 4 in subjects with vIGA-AD score 'moderate' at randomization. The same testing method, the stratified CMH test, will be used as for the primary endpoint.

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 application (including unscheduled/retest assessments). The randomization date will be used in place of the date of first treatment application for randomized subjects who discontinue without having treatment applied. 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.

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 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 and Physical Examination

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)
(-6, 1) *	Baseline
(2, 8)	Week 1
(9, 15)	Week 2
(16, 22)	Week 3
(23, 29)	Week 4

* 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 success = 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 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 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 the analysis visit window or is prior to the start of the analysis visit window.

- EASI total score = $0.2\ (E_h + I_h + E_{Xh} + L_h)\ A_h + 0.2\ (E_u + I_u + E_{Xu} + L_u)\ A_u + 0.3\ (E_t + I_t + E_{Xt} + L_t)\ A_t + 0.3\ (E_l + I_l + E_{Xl} + L_l)\ A_l$

  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.

- 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.

- 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.
- IDQoL Score = sum of the 10 questions (ranging from 0 to 3), resulting in a score ranging from 0 to 30. The higher the score, the more quality of life is impaired. If 1 item is missing, that item is scored as 0. If 2 or more items are missing, the score should not be calculated. The severity of dermatitis is scored separately, and it is not included in the IDQoL total score.
- 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.
- 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.

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 0.025. Confidence intervals (CIs) will be two-sided with both 97.5% and 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

The covariate for this study is baseline vIGA-AD (vIGA-AD=2 - Mild vs. vIGA-AD=3 - Moderate at randomization). The CMH test will not be stratified by study site as the 650 subjects will be randomized by over 90 sites leading to a low subject to site ratio. 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. This procedure will ensure that the data collected on or after these specific 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, 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 datasets, the necessary derived variable will be computed as defined in Section 5.5 and analyzed using a CMH analysis, adjusted for the randomized vIGA-AD score for the primary efficacy endpoint. The treatment effect on the secondary efficacy endpoint dealing with vIGA-AD success in the vIGA-AD moderate ITT population will be assessed with an unstratified chi-square test. Results will be combined into one multiple imputation inference as follows:
    a. Common proportion of success, common Mantel-Haenszel (MH) proportion difference (and associated 97.5% and 95% CIs), and common MH odds ratio (and associated common 97.5% and 95% CIs) 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 97.5% and 95% CIs 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.
   ii. 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 CIs 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 EASI total score (refer to Section 5.5) where the missing EASI total score data at Weeks 1, 2, and 4 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.

Even though it is only considered an exploratory endpoint, the multiple imputation methods will also be applied to the weekly average WI-NRS score at Weeks 1, 2, 3, and 4. 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 average weekly WI-NRS score will include all subjects in 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 score 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), 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 EASI-75 where subjects with missing baseline EASI total score will be considered as non-responder.

Subjects who discontinued early from study due to an AE or lack of efficacy will be considered as a non-responder 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.

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, 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., $\alpha \leq 0.025$) to non-significant (i.e., $\alpha > 0.025$) 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.

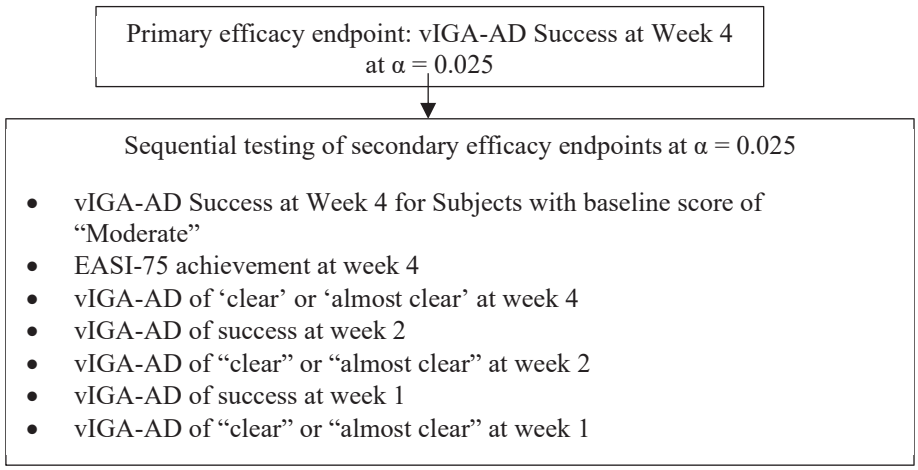
6.4 Multicenter Studies

The study will be conducted at approximately 111 study sites in the US, Canada and Poland. During the conduct of the study, additional countries and/or sites may be added if necessary. Given the large number of sites, analyses will not be stratified by study site.

6.5 Multiple Comparisons/Multiplicity

To control for familywise type I error at level of 0.025, the secondary efficacy endpoints will be tested sequentially at an alpha level of 0.025 upon successful demonstration of statistical significance for the primary endpoint. The order of the secondary efficacy endpoints is important. Should, anywhere during the sequential testing of secondary efficacy endpoints, there is a p-value >0.025, further testing results are considered "not statistically significant" regardless of the p-value computed.

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.

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, secondary and WI-NRS success efficacy endpoints:

- 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 groups 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:

$$\text{Number of days in study} = \text{Date of completion/discontinuation} - 1^{st} \text{ dose date}^{*} + 1$$

* 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 ITT 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)
- Sex at birth
- Ethnicity
- Race*
- Baseline Height (cm)
- Baseline Weight (kg)
- Baseline Body Mass Index (BMI) percentile
- Fitzpatrick Skin Type
- Prior failure of Topical Corticosteroids, Topical Calcineurin Inhibitors, Eucrisa
- Atopic Dermatitis involvement on the face, on the eyelids

- Number of AD flares in the last 12 months
- Number of prior prescription treatments ever used
- Duration of disease (months)
- 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 CDLQI (≥ 4 years of age)
- Baseline IDQoL (< 4 years of age)
- Baseline DFI
- Baseline POEM

*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.

BMI (kg/m^2) = (weight in kg)/ [(height in $cm/100)^2$]. Baseline height will be used to derive BMI for all post-baseline visits until height is collected again (planned at week 4). After BMI is calculated using the same formula above for children, 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).

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 or later.

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 March 2022 or later.

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 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 0.05%, 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 0.05% 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 0.05% and matching vehicle cream. The alternative hypothesis is that the vIGA-AD success does differ between roflumilast cream 0.05% 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 0.05%
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 both 97.5% and 95% Wilson CIs. The common MH odds ratio and common MH proportion difference, adjusted for the randomization factor (i.e., randomized vIGA-AD score) will also be provided along with their common associated 97.5% and 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 randomized baseline vIGA-AD. 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 2.5% 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 randomized baseline vIGA-AD, for the common MH proportion difference at Week 4.

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

- 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 4.2).

For the last three sensitivity/supplemental 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 individual study sites on the primary analysis of the primary endpoint, the following analyses will be performed:

- To assess the impact of site on the primary analysis endpoint, the proportion of subjects achieving vIGA-AD success and 97.5% CI within each site with at least 10 randomized subjects will be tabulated. No p-values will be provided. Forest plots of the proportions (and associated 97.5% CI) for each site with at least 10 randomized subjects, by 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 sites that randomized at least 10 subjects. To do so, the primary analysis of the primary efficacy endpoint will be repeated by removing a different study site for each iteration. Forest plots of the odds ratio (and associated 97.5% CI) for each analysis will also be provided.

12.2 Secondary Efficacy Endpoints Analysis

Secondary efficacy endpoints are based on the vIGA-AD 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.
- 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 endpoints will be tested sequentially at an alpha level of 0.025.

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 4.3) and use of a chi-square test to evaluate the treatment effect. 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 used.

The remaining secondary efficacy endpoints, 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, vIGA-AD of 'clear' or 'almost clear' at Week 2, vIGA-AD of success at Week 1 and vIGA-AD of 'clear' or 'almost clear' at Week 1 will be performed as described for the primary efficacy endpoint (refer to Section 12.1.3) based on the ITT population (refer to Section 4.1).

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:

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

For these 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:

- Non-responder imputation on ITT population.

- Observed data on ITT population.

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

12.3 Exploratory Efficacy Endpoints Analysis

Exploratory efficacy endpoints are based on the vIGA-AD scale, EASI questionnaire, WI-NRS scale, %BSA affected by AD, CDLQI/IDQoL questionnaires, DFI questionnaire, SCORAD tool or POEM tool. The list of other 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 EASI questionnaire, refer to Section 12.2.
- For more details about the WI-NRS scale, refer to Section 5.5.
- 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 is a self-administered validated questionnaire designed to measure the health-related quality of life of children subjects (≥4 years old) 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 question 7, if last week was a school time, question 7 is rated from 0 (Not at all) to 3 (Prevented school) but if the last week was a holiday time, question 7 is rated from 0 (Not at all) to 3 (Very much).
- The IDQoL questionnaire is designed to assess the impact of atopic dermatitis on the quality of life of infants <4 years old. It is self-explanatory and should be completed by the child's parent(s) or regular caregiver and consists of 10 questions rated from 0 to 3. The severity of dermatitis is scored separately, and it is not included in the IDQoL total score.
- 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 parents/caregivers 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 subjects who discontinue due to lack of efficacy or adverse event, efficacy data assigned to a pre-specified analysis visit will be removed from observed data 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. This will be true for all endpoints described above.

For the continuous exploratory efficacy endpoints of change and percent change from baseline in average weekly WI-NRS, EASI total score, % BSA, CDLQI/IDQoL score (+ severity of dermatitis), DFI score, SCORAD score, and POEM score at Week 1, Week 2, Week 3 (WI-NRS only), 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. For the continuous other efficacy endpoint of daily WI-NRS, descriptive statistics will also be presented similarly on a daily 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.

Continuous exploratory efficacy endpoints will be analyzed at Week 1, Week 2, Week 3 (WI-NRS only), and Week 4 and for daily WI-NRS assessments using an analysis of covariance (ANCOVA) with the factors of treatment, randomized vIGA-AD score, and baseline of the variable under analysis as covariate. 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 97.5% and 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.

Exploratory categorical endpoints based on vIGA-AD, EASI, or WI-NRS, ie VIGA-AD score of clear, EASI-50, EASI-75, EASI-90, EASI-100, WI-NRS success in the WI-NRS population, will

be analyzed by visit and treatment group based on MI data and the ITT population with the same methods used for related primary or secondary endpoints. For all categorical exploratory efficacy endpoints, count and percentages of subjects meeting the criteria for exploratory efficacy endpoints will be presented by visit and treatment group based on observed data in the ITT population along with a 97.5% and 95% Wilson CI. The observed data will also be analyzed using a CMH test adjusted for randomized vIGA-AD score. Common MH odds ratio, common MH proportion difference and their common associated 97.5% and 95% CIs, adjusted for the randomized vIGA-AD score, will be provided. The p-value will be from a CMH test, 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 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) and WI-NRS success will be repeated by subgroups (refer to Section 6.6) based on the multiple imputation data using the ITT population (primary and secondary endpoints) and the WI-NRS population (WI-NRS success). 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 odds ratios and difference in percent of subgroup analysis for primary efficacy endpoint, secondary efficacy endpoints and WI-NRS success will also be provided.

No subgroup analyses will be presented for the exploratory efficacy endpoint (refer to Section 2) except for WI-NRS success.

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 5 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 ITT, Tipping point (CMH)) #2 ITT, non-responder imputation (CMH) #3 ITT, observed data (CMH) #4 PP, observed data (CMH)
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 (chi-square)	#1 vIGA-AD Moderate ITT, non-responder imputation (chi-square) #2 vIGA-AD Moderate ITT, observed data (chi-square)
-EASI-75 at week 4 -vIGA-AD score of '0' or '1' at week 4 -vIGA-AD Success at week 2 -vIGA-AD score of '0' or '1' at week 2 -vIGA-AD Success at week 1 -vIGA-AD score of '0' or '1' at week 1	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.

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. 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, TEAE leading to discontinuation of study drug, 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 and hematology, change from baseline and percent change from baseline values will be presented by treatment group at each scheduled visit.

Shift tables from baseline to each post-baseline assessments describing shifts to out-of-normal range will be provided for chemistry and hematology. 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. The BMI percentile rather than BMI kg/m^2 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.

A listing of all vital sign assessments including weight, BMI, and BMI percentile 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 parent/caregiver for the subject will be summarized using number and percentage similarly.

13.5 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

PK data will be included in SDTM/ADaM datasets. A PK report will be provided by the study Pharmacokineticist.

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 0.05%	Roflumilast Cream 0.05%
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.