1.0 Title Page

Statistical Analysis Plan

Study M16-048

Phase 2b

A Phase 2b Multicenter, Randomized, Placebo Controlled, Double-Blind Dose-Ranging Study to Evaluate ABT-494 (Upadacitinib) in Adult Subjects with Moderate to Severe Atopic Dermatitis

Date: 23 Aug 2017

Version 1.0

2.0 Table of Contents

List of Tables

List of Figures

List of Appendices

3.0 Introduction

This statistical analysis plan (SAP) describes the statistical analysis to be completed by the AbbVie Data and Statistical Science Department for Upadacitinib study Protocol M16-048 dated 07 June 2017, which incorporates four amendments (original protocol: 24 June 2016; Amendment 1: 25 July 2016; Amendment 2: 18 October 2016; Amendment 3: 16 December 2016; Amendment 4: 07 June 2017). It provides details to further elaborate statistical methods as outlined in the protocol and describes analysis conventions to guide the statistical programming work.

This is the first version of the SAP for Protocol M16-048.

Unless noted otherwise, all analyses will be performed using SAS version 9.2 or later (SAS Institute Inc., Cary, NC 27513) under the UNIX operating system.

4.0 Study Objectives, Design and Procedures

4.1 Objectives

The primary objective of this study is to evaluate the efficacy of multiple doses of upadacitinib compared with placebo in the treatment of adults with moderate to severe atopic dermatitis (AD). The secondary objectives of this study are to evaluate the safety of upadacitinib compared with placebo as measured by adverse events (AEs), laboratory data, vital signs and Electrocardiogram (ECG).

4.2 Design Diagram

This is an 88-week Phase 2b, randomized, double-blind, parallel-group, placebo-controlled multicenter study of safety and efficacy of upadacitinib in adult subjects with moderate to severe AD.

The duration of the study will be up to 88 weeks and will include a 35-day maximum screening period, a 16-week double-blind treatment period (Period 1) and a 72-week double-blind treatment period (Period 2). In addition, a 30-day follow-up visit (or phone

call if a visit is not possible) should occur to determine the status of any ongoing AEs or serious adverse events (SAEs), or the occurrence of any new AEs/SAEs.

Period 1: A 16-week double-blind, placebo-controlled treatment period during which subjects will be randomized in a 1:1:1:1 ratio to one of four treatment groups:

- Group 1: upadacitinib 7.5 mg once daily (QD) for 16 weeks
- Group 2: upadacitinib 15 mg QD for 16 weeks
- Group 3: upadacitinib 30 mg QD for 16 weeks
- Group 4: Matching Placebo for 16 weeks

Randomization will be stratified by geographic region (US/PR/Canada, EU/AUS, and Japan). Enrollment in Japan will be capped at 10% of subjects (4 subjects per group).

In Period 1, discontinuation from study drug will be mandatory for any subject with an Eczema Area and Severity Index (EASI) score worsening of 25% or more compared with their baseline EASI score at any 2 consecutive scheduled study visits from Week 4 to Week 12.

Period 2: a 72-week double-blind, placebo controlled treatment period during which subjects who complete Period 1 will be re-randomized at Week 16 within their original treatment group assignments to either upadacitinib or placebo into a 72-week double-blind, placebo controlled treatment period in a 1:1 ratio as shown in Figure 1 (study design). Randomization will be stratified by geographic region and the improvement of EASI score by at least 75% (EASI 75 response) at Week 16. At the Week 16 visit, all subjects will be re-randomized as follows into Period 2:

- Group 1 in Period 1: upadacitinib 7.5 mg QD or matching placebo for 72 weeks
- Group 2 in Period 1: upadacitinib 15 mg QD or matching placebo for 72 weeks
- Group 3 in Period 1: upadacitinib 30 mg QD or matching placebo for 72 weeks

- Group 4 in Period 1: upadacitinib 30 mg QD or matching placebo for 72 weeks

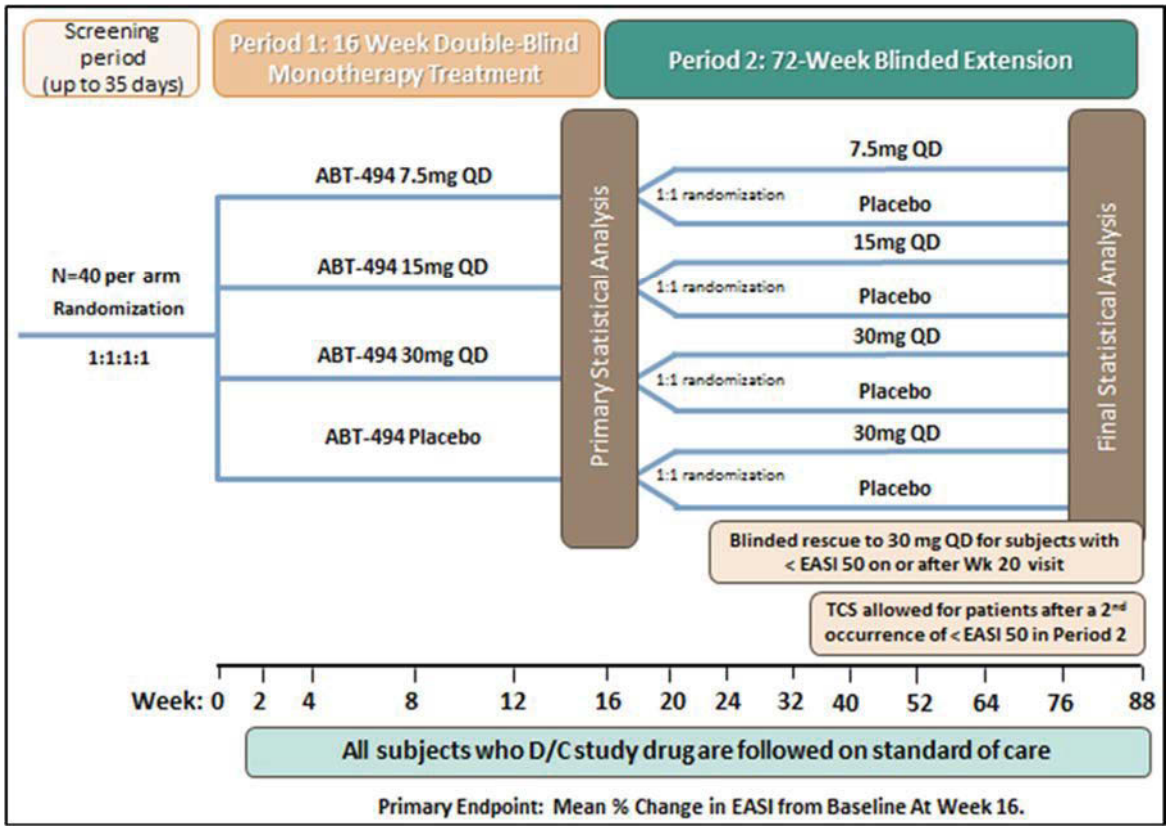
Starting at the Week 20 visit (4 weeks after re-randomization into Period 2), rescue therapy with upadacitinib 30 mg QD will be provided to subjects who have a first instance of an EASI score improvement of less than 50% (i.e., < EASI 50 response), when compared against the Baseline. These subjects will continue on upadacitinib 30 mg QD for the remainder of the study.

During Period 2, concomitant class III – IV, medium potency topical corticosteroid treatment will be permitted at the Week 24 visit and thereafter in subjects after a second instance of < EASI 50 response in any two Period 2 study visits beginning from Week 20 (see Protocol Section 5.2.3.4 for permitted treatments). For subjects who receive topical corticosteroid rescue therapy, a visit will be required 4 weeks after receiving the rescue therapy. Discontinuation from study drug will be mandatory for subjects with < EASI 50 response compared with their baseline EASI score 4 weeks following rescue with topical corticosteroids or at any visit thereafter.

In order to minimize missing data for efficacy and safety assessments, subjects who prematurely discontinue study drug should continue to be followed for all regularly scheduled visits, unless they have decided to discontinue the study participation entirely (withdrawal of informed consent).

A schematic of the overall study design is shown in Figure 1.

Figure 1. Study Design Schematic



4.3 Sample Size

Approximately 160 subjects will be randomized to the four treatment groups in a ratio of 1:1:1:1. The sample size for this study is based on the percent change in EASI from baseline at Week 16. Assuming a percent change in EASI from baseline at Week 16 of 35, 45, 60, and 70 in the placebo, 7.5 mg, 15 mg, and 30 mg arms with a standard deviation 40 and a maximum efficacy of 80, a sample size of 40 subjects per treatment group is sufficient to test for the presence of a dose response signal, to select the best dose response model for the observed data out of a pre-specified set of candidate models, and to estimate target doses of interest (e.g., the minimum effective dose, MED) via modeling

using the MCP-Mod (Multiple comparison procedure and modeling) approach. This approach provides 99% average power to detect a dose effect at the 5% level of significance (one-sided) with the linear, E_{max}, exponential, logistic and sigE_{max} models pre-specified as likely candidates to characterize the dose-response for upadacitinib for the percent change in EASI.

A sample of size 40 per group provides 97% power to detect a significant difference between 30 mg QD and placebo, and 78% power to detect a significant difference between 15 mg QD and placebo at the two-sided level of significance of 5.0%.

4.4 Interim Analysis

The analysis of safety and efficacy data for the first double blind period will be conducted after all subjects have either completed Period 1 or discontinued from the study. This efficacy analysis is the only and final analysis of Period 1, thus no adjustment of alpha-level is needed. In addition, there will be an interim analysis after all subjects have either reached Week 32 or discontinued from the study.

An external Data Monitoring Committee (DMC) comprised of persons independent of AbbVie and with relevant expertise in their field will review unblinded safety data from the ongoing study. If necessary the DMC will also be given access to efficacy data upon request. A DMC charter was prepared to describe the roles and responsibilities of the DMC members, frequency of data reviews, and relevant safety data to be assessed.

Communications from the DMC to the Study Teams will not contain information that could potentially unblind the team to subject treatment assignments.

5.0 Analysis Populations

5.1 Analysis Populations

Intent-to-Treat (ITT) Population

An ITT population will be defined for each period and be used for the efficacy analyses.

- The ITT Population in Period 1 (ITT_1) is defined as all subjects who are randomized at Day 1.
- The ITT Population in Period 2 (ITT_2) is defined as all subjects who are re-randomized at the entry of Period 2.

For efficacy analyses, subjects included in the ITT populations will be analyzed as randomized. If a subject is randomized or re-randomized in the wrong strata, the subject will be analyzed according to the actual strata that subject belongs to.

Safety Population

A Safety population will be defined for each period and be used for safety analyses.

- The Safety Population in Period 1 (Safety_1) is defined as all subjects who are randomized and receive at least one dose of study drug in Period 1.
- The Safety Population in Period 2 (Safety_2) is defined as all subjects who are re-randomized at the entry of Period 2 and receive at least one dose of study drug in Period 2.

For the Safety populations, subjects are assigned to a treatment group based on the "as treated" treatment group. In the case where the subject received more than one treatment, the "as treated" is determined by the dosage that a patient received the most during a period.

Table 1. Notations for Treatment Groups

Population	Treatment Code	Definition
ITT_1 and Safety 1	UPA 7.5 MG	Upadacitinib 7.5 mg.
	UPA 15 MG	Upadacitinib 15 mg.
	UPA 30 MG	Upadacitinib 30 mg.
	PBO	Placebo
ITT_2 and Safety 2	UPA 7.5 MG/UPA 7.5 MG	Upadacitinib 7.5 mg in Period 1 and upadacitinib 7.5 mg in Period 2
	UPA 7.5 MG/PBO	Upadacitinib 7.5 mg in Period 1 and placebo in Period 2
	UPA 15 MG/UPA 15 MG	Upadacitinib 15 mg in Period 1 and upadacitinib 15 mg in Period 2
	UPA 15 MG/PBO	Upadacitinib 15 mg in Period 1 and placebo in Period 2
	UPA 30 MG/UPA 30 MG	Upadacitinib 30 mg in Period 1 and upadacitinib 30 mg in Period 2
	UPA 30 MG/PBO	Upadacitinib 30 mg in Period 1 and placebo in Period 2
	PBO/UPA 30 MG	Placebo in Period 1 and upadacitinib 30 mg in Period 2
	PBO/PBO	Placebo in Period 1 and placebo in Period 2

5.2 Variables Used for Stratification of Randomization

For Period 1, the randomization will be stratified by geographic region (US/PR/Canada, EU/AUS, and Japan); for Period 2, the re-randomization will be stratified by geographic region and EASI 75 response at Week 16.

6.0 Analysis Windows

Definition of Baseline

Baseline refers to the last non-missing observation prior to the first dose of study drug. For randomized but not treated subjects, the baseline value will be the last non-missing measurement recorded prior to the randomization.

Zero or Missing Baseline Values when Calculating Percent Change from Baseline

When baseline value is missing for a subject, percent change from baseline cannot be calculated and therefore percent change values will be set to missing in these situations. When the baseline value is 0, the percent change will be set to missing, except when the post baseline value is also 0, in which case the percent change will be set to 0.

Definition of Final Observation for Safety Analysis

Period 1: The final observation is defined as the last non-missing observation collected

- within 30 days following the last dose of study drug in Period 1 for subjects who discontinued from the study drug without having any dose of study drug in Period 2,
- on or before the day of the first dose of study drug in Period 2 for subjects who had at least one dose of the study drug in Period 2.

Laboratory and vital sign evaluations performed on the day of the first dose of study drug in Period 2 will be included in the analysis for Period 1. AEs with onset on the date of the first dose of Period 2 will be attributed to Period 1.

Period 2: The final Observation in each period is defined as the last non-missing observation collected within 30 days following the last dose of study drug in the respective period.

Definition of Rx Days (Days Relative to the First Dose of Study Drug)

Rx days are calculated for each time point relative to the date of first dose of study drug. They are defined as the number of days between the day of the first dose of study drug and the specific time point. Rx days are negative values when the time point of interest is prior to the first study drug dose day. Rx days are positive values when the time point of interest is on or after the first study drug dose day. The day of the first dose of study drug is defined as Rx Day 1, while the day prior to the first study drug dose is defined as Rx Day –1 (there is no Rx Day 0).

For analysis in Period 1 (P1 Rx Day) and Period 2 (P2 Rx Day), the Rx Day is calculated relative to the first dose of study drug in Period 1 and Period 2, respectively.

For subjects randomized but not dosed, Rx Days are calculated relative to the randomization.

Definition of Analysis Windows

All time points and corresponding time windows in Periods 1 and 2 are defined based on P1 Rx Days P1 and P2 Rx Day, respectively.

Analysis time windows are constructed using the following algorithm:

- Determine the nominal Rx Day for each scheduled visit (e.g., Week 4 [4 weeks after baseline visit] equals P1 Rx Day 29).
- In order to include all post baseline data, the first post-baseline interval starts on the first day after the first dose of study drug (P1 or P2 Rx Day 2).
- Determine the window around a specific nominal Rx Day by adding or subtracting half of the interval between adjacent visits (e.g., days between Week 2 and Week 4 is 14). The threshold between adjacent visits is determined by splitting the interval evenly between the visits. If the resulting split is between Rx days, then the threshold is determined as the midpoint between the adjacent visits. If the resulting split is on an Rx day, then the threshold is determined as being between that Rx day and the Rx day prior to it (e.g., the split between Week 2 and Week 4 would be between P1 Rx Days 22 and 23).
- If more than one assessment is included in a time window, the assessment closest to the nominal day will be used. If there are two observations equidistant to the nominal day, the one after the nominal day will be used in analyses. If more than one assessment is included on the same day, then the worst assessment on that day will be used in analyses, except otherwise specified.

For subjects randomized but not dosed, Rx Day calculated relative to Randomization in each period.

For the efficacy variables, vital signs, and laboratory values, visit windows are defined as follows.

Table 2. Analysis Windows for EASI, SCORAD, IGA, BSA, PGIS, Laboratory Parameters, and Vital Sign Variables (Period 1)

Scheduled Week	Nominal Day	P1 Rx Day Interval
Baseline	1	≤ 1
Week 2	15	2 – 22
Week 4	29	23 – 43
Week 8	57	44 – 71
Week 12	85	72 – 99
Week 16	113	100 – 126[a]

a. Minimum of the upper bound and the day of the first dose in Period 2.

Table 3. Analysis Windows for Pruritus NRS, ADerm-SS, ADerm-IS, and Actigraphy (Period 1)

Scheduled Week	Nominal Day	P1 Rx Day Interval
Baseline	–1	< 1[a]
Week 1	8	2 – 11
Week 2	15	12 – 18
Week 3	22	19 – 25
Week 4	29	26 – 32
Week 5	36	33 – 39
Week 6	43	40 – 46
Week 7	50	47 –53
Week 8	57	54 – 60
Week 9	64	61 – 67
Week 10	71	68 – 74
Week 11	78	75 – 81
Week 12	85	82 – 88
Week 13	92	89 – 95
Week 14	99	96 – 102
Week 15	106	103 – 109
Week 16	113	110 – 116[b]

a. ADerm-IS CCI [redacted] assessments collected on or before Day 1 will be used as baseline (P1 Rx Day Interval: ≤ 1).

b. Minimum of the upper bound and the day of the first dose in Period 2 for efficacy analyses.

Table 4. Analysis Windows for DLQI, MOS, and POEM (Period 1)

Scheduled Week	Nominal Day	P1 Rx Day Interval
Baseline	1	≤ 1
Week 4	29	2 – 71
Week 16	113	72 – 154[a]

a. Minimum of the upper bound and the day of the first dose in Period 2.

Table 5. Analysis Windows for Asthma Symptoms Questionnaire and Daytime Nasal Symptoms Questionnaire (Period 1)

Scheduled Week	Nominal Day	P1 Rx Day Interval
Baseline	1	≤ 1
Week 16	113	2 – 224[a]

a. Minimum of the upper bound and the day of the first dose in Period 2.

Table 6. Analysis Windows for Blood Sample Parameters* (Period 1)

Scheduled Week	Nominal Day	P1 Rx Day Interval
Baseline	1	≤ 1
Week 8	57	2 – 85
Week 16	113	86 – 140[a]

* Including T (CD4+ and CD8+) cells, B (CD19+) cells, natural killer (NK) cells, and natural killer-T (NKT) cells.

a. Minimum of the upper bound and the day of the first dose in Period 2.

Table 7. Analysis Windows for EASI, SCORAD, IGA, BSA, PGIS, Laboratory Parameters and Vital Sign Variables (Period 2)

Scheduled Week	Nominal Day	P2 Rx Day Interval
Entry of Period 2 (Week 16)	1	≤ 1
Week 20	29	2 – 43
Week 24	57	44 – 85
Week 32	113	86 – 141
Week 40	169	156 – 211
Week 52	253	212 – 295
Week 64	337	296 – 379
Week 76	421	378 – 463
Week 88	505	464 – 546

Table 8. Analysis Windows for Daily Pruritus NRS, ADerm-SS, and ADerm-IS (Period 2)

Scheduled Week	Nominal Day	P2 Rx Day Interval
Entry of Period 2 (Week 16)	1	< 1[a]
Week 17	8	2 – 11
Week 18	15	12 – 18
Week 19	22	19 – 25
Week 20	29	26 – 32
Week 21	36	33 – 39
Week 22	43	40 – 46
Week 23	50	47 – 53
Week 24	57	54 – 60
Week 25	64	61 – 67
Week 26	71	68 – 74
Week 27	78	75 – 81
Week 28	85	82 – 88
Week 29	92	89 – 95
Week 30	99	96 – 102
Week 31	106	103 – 109
Week 32	113	110 – 116
Week 33	120	117 – 123
Week 34	127	124 – 130
Week 35	134	131 – 137
Week 36	141	138 – 144
Week 37	148	145 – 151
Week 38	155	152 – 158
Week 39	162	159 – 165
Week 40	169	166 – 211
Week 52	253	212 – 295
Week 64	337	296 – 379
Week 76	421	380 – 463
Week 88	505	464 – 546

a. ADerm-IS[redacted] assessments collected on or before Day 1 of Period 2 will be used as entry of Period 2 (P2 Rx Day Interval: ≤ 1).

Table 9. Analysis Windows for DLQI (Period 2)

Scheduled Week	Nominal Day	P2 Rx Day Interval
Entry of Period 2 (Week 16)	1	≤ 1
Week 40	169	2 – 211
Week 52	253	212 – 295
Week 64	337	296 – 379
Week 76	421	378 – 463
Week 88	505	464 – 546

Table 10. Analysis Windows for MOS, and POEM (Period 2)

Scheduled Week	Nominal Day	P2 Rx Day Interval
Entry of Period 2 (Week 16)	1	≤ 1
Week 24	57	2 – 113
Week 40	169	114 – 211
Week 52	253	212 – 295
Week 64	337	296 – 379
Week 76	421	378 – 463
Week 88	505	464 – 546

Table 11. Analysis Windows for Asthma Symptoms Questionnaire and Daytime Nasal Symptoms Questionnaire (Period 2)

Scheduled Week	Nominal Day	P2 Rx Day Interval
Entry of Period 2 (Week 16)	1	≤ 1
Week 40	169	2 – 336

Table 12. Analysis Windows for Blood Sample Parameters* (Period 2)

Scheduled Week	Nominal Day	P2 Rx Day Interval
Entry of Period 2 (Week 16)	1	≤ 1
Week 20	29	2 – 56

* Including T (CD4+ and CD8+) cells, B (CD19+) cells, natural killer (NK) cells, and natural killer-T (NKT) cells.

Dealing with Multiple Measurements Collected on the Same Day

For efficacy related analyses, if there are multiple measurements for a particular parameter collected on the same day for the same subject, the most conservative measurement indicating the worst disease status will be used in the analyses.

For safety related analyses, if there are multiple measurements made for a particular laboratory or vital sign parameter on the same day for the same subject without the actual times, the average of the values will be used in the mean change analyses. For summaries and listings for shift from baseline on the minimum, maximum, final, and potentially significant values, all collected values within the defined treatment-emergent window will be used.

Definition of Missing Data Imputation

No global imputation is taking place at the database level. Efficacy related imputations are outlined in Section 10.1.5. There is no imputation for missing values in the safety analyses.

7.0 Demographics, Baseline Characteristics, Medical History, and Previous/Concomitant Medications

7.1 Demographic and Baseline Characteristics

Demographics and Baseline characteristics will be summarized for each arm and for overall of the ITT populations. Continuous variables will be summarized with the number of non-missing observations by mean, standard deviation, first quartile, median, third quartile, minimum and maximum values. Categorical data will be summarized using frequencies and percentages.

The following demographic and baseline parameters will be summarized.

Subject Demographics

- Sex (Male, Female)

- Age (years), defined as the number of years from year of birth to year of first dose of study drug
- Age categories (< 40 years, 40 – 65 years, ≥ 65 years)
- Race (White, Black, American Indian/Alaska Native, Native Hawaiian or Other Pacific Islander, Asian, Other)
- Ethnicity (Hispanic or Latino, not Hispanic or Latino)
- Body weight (kg)
- Height (cm)
- BMI (kg/m^2)
- BMI category (< 25, ≥ 25 – < 30, ≥ 30 – < 40, ≥ 40)
- Geographic Region (US/PR/Canada, EU/AUS, Japan)

Tobacco/Nicotine and Alcohol Use

- Tobacco/Nicotine Use (unknown, never, current, former)
- Alcohol Use (unknown, never, current, former)

Baseline Characteristics

- EASI overall score and body region scores
- Investigators Global Assessment (IGA)
- Body Surface Area (BSA) in percentage
- Scoring Atopic Dermatitis (SCORAD)
- Dermatology Life Quality Index (DLQI)
- Daily pruritus Numerical Rating Scale (NRS)
- Patient Global Impression of Severity (PGIS)
- Atopic Dermatitis Symptom Scale (ADerm-SS)
- Atopic Dermatitis Impact Scale (ADerm-IS)
- Medical Outcomes Study (MOS) Sleep Scale
- Patient Oriented Eczema Measure (POEM)
- Asthma Symptoms and Daytime Nasal Scores (subjects with symptoms)

- Actigraphy data: total sleep time per night (TST min), Sleep Efficiency (%), Wake After Sleep Onset (WASO), number of scratching events per hour, mean activity during rest (sleep)
- Disease duration since diagnosis (years)
- Disease duration since symptoms started (years)
- Duration between symptoms and diagnosis (years)
- Prior Atopic Dermatitis Treatment
- Electrocardiogram (ECG), containing categories including normal, abnormal – not clinically significant, abnormal – clinically significant, unable to evaluate, and details of the abnormality when applicable.
- TB Status: Tuberculin PPD skin test result, QuantiFERON-TB Gold test result

- Chest x-ray
  - Normal, Abnormal
  - Calcified granulomas (Absent, Present)
  - Pleural scarring (Absent, Present)
  - Pleural thickening (Absent, Present)

7.2 Medical History

For the ITT_1 population, the medical history data will be summarized and presented using body systems and conditions/diagnoses as captured on the CRF. The body systems will be presented in alphabetical order and the conditions/diagnoses will be presented in alphabetical order within each body system. The number and percentage of subjects with a particular condition/diagnosis will be summarized for each treatment group. Subjects reporting more than one condition/diagnosis within a body system will be counted only once for that body system. If no subject reported information for a specific body system, then that category will not be displayed in the summary table.

7.3 Previous Treatment and Concomitant Medications

Prior and concomitant medications will be summarized by generic name. A prior medication is defined as any medication taken prior to the first dose of study drug. A

concomitant medication is defined as any medication that started prior to the first dose of study drug and continued to be taken after the first dose of study drug or any medication that started after the first dose of study drug, but not after the last dose of study drug. The number and percentage of subjects who had taken medications will be summarized by generic drug name assigned by the World Health Organization (WHO) for both prior and concomitant medications.

7.4 Protocol Deviations

Number and percentage of subjects who reported at least one of the following protocol deviation categories will be provided.

- Subject entered into the study even though she/he did not satisfy entry criteria
- Subject who developed withdrawal criteria during the study and was not withdrawn
- Subject who received wrong treatment or incorrect dose
- Subject who received excluded or prohibited concomitant treatment

8.0 Patient Disposition

The number of subjects for each of the following categories will be summarized, for overall and for each treatment group in the ITT_1 Population:

- Number of subjects randomized to Period 1
- Number of subjects treated in Period 1
- Number of subjects who completed Period 1
- Number of subjects who discontinued study drug during Period 1
- Number of subjects who prematurely discontinued from Period 1

The number of subjects for each of the following categories will be summarized, for overall and for each treatment group in the ITT_2 Population:

- Number of subjects re-randomized to Period 2

- Number of subjects treated in Period 2
- Number of subjects who completed Period 2
- Number of subjects who discontinued study drug during Part Period 2
- Number of subjects who prematurely discontinued from Period 2

In addition, the reasons (primary and all reason) for premature discontinuation of study drug and discontinuation from the study will be summarized with frequencies and percentages.

9.0 Study Drug Exposure and Compliance

9.1 Study Drug Exposure

Summary of study drug exposure and study drug compliance will be provided for each treatment arm for ITT_1 and ITT_2.

Study drug exposure (days) will be summarized using the sample size, mean, standard deviation, minimum, median and maximum for each treatment part. Study drug exposure will be summarized as follows:

Period 1:

For subjects who did not continue into Period 2:

- Date of last dose of study drug in Period 1 – Date of first dose of study drug in Period 1 + 1.

For subjects who continued into Period 2: the minimum of

- Date of first dose of study drug in Period 2 – Date of first dose of study drug in Period 1.
- Date of last dose of study drug in Period 1 – Date of first dose of study drug in Period 1 + 1.

Period 2:

- Date of last dose of study drug in Period 2– Date of first dose of study drug in Period 2 + 1.
- For interim analysis at Week 32, minimum of the above and (cutoff date – Date of first dose of study drug in Period 2 + 1).

Additional summaries of upadacitinib exposures by dose level will also be performed including exposure to each dose level across periods.

9.2 Compliance

Treatment compliance (TC) will summarized for each treatment group in each period. The treatment compliance is defined as the number of tablets actually taken by the subject divided by the number of tablets planned to be taken by the subject during the double-blind treatment period of the study. Of note, the compliance in Period 2 will be summarized up to the date prior to the first dose of rescue. Based on a daily dose of one tablet for the study drug, TC will be calculated using the following formula:

$$TC = \frac{\text{number of tablets taken}}{\text{number of days from first to last dose of blinded study drug}} \cdot 100\%$$

10.0 Efficacy Analysis

10.1 General Considerations

Efficacy analyses will be carried out in ITT_1 and ITT_2, respectively for Periods 1 and 2. Pairwise comparisons of each upadacitinib treatment group versus placebo will be performed in Period 1. Descriptive summaries will be provided for all treatment groups in Period 2.

10.1.1 Analysis of Binary Variables

For binary variables, frequencies and percentages will be reported for each treatment group. Pairwise comparison of each upadacitinib group and placebo will be performed using the Cochran Mantel-Haenszel test adjusting for stratification factors. If there is a stratum for a treatment group that has no subject in it, a value of 0.1 will be added to all cells in the corresponding table in order to prevent dividing by 0, as suggested in Greenland and Robins (1985).[1] Point estimates and 95% CIs for the difference in proportions between each of upadacitinib treatment groups and placebo will be provided. Construction of CIs for the common risk difference will be based on the Mantel-Haenszel estimate adjusting for stratification factors.

10.1.2 Analysis of Continuous Variables

For continuous variables, the model based mean and standard error will be provided. The Baseline and visit means will also be presented for each treatment group. The treatment groups will be compared using the analysis of covariance (ANCOVA) model with treatment group and stratification factors as fixed effects, and the corresponding baseline value as covariates. Point estimates and 95% CIs of mean change from baseline within treatment groups, and between each upadacitinib treatment group and placebo will be provided.

10.1.3 Analysis of Time to Event Variables

For the time to event endpoints, the time to event will be calculated as:

- Time to first event (as observed) = [date of first event in Period 1] – [date of first study drug in Period 1] + 1;
- If a subject never attains an event in Period 1, then that subject's time to first achievement will be censored at the last visit where the variable was measured in Period 1.

Time to EASI 50/75/90 response and IGA "0" or "1" response will be evaluated for Period 1. Time to loss of EASI 50 response among those who were re-randomized as EASI 75 responders at Week 16 will be evaluated for Period 2.

Time to event variables will be analyzed by stratified log-rank test adjusting for stratification factors.

10.1.4 Analysis of Dose Response

The dose-response relationship among the 3 upadacitinib dose groups and the placebo group will be characterized for the primary efficacy endpoint percent change from Baseline in EASI score at Week 16 and two secondary endpoints of EASI 75 and IGA 0/1 at Week 16 using the Multiple Comparison Procedure – Modeling (MCP-Mod)[2,3] method. The response based on the LOCF imputation will be used, and ADDPLAN DF software will be used to perform the MCP-Mod analyses.

A set of six pre-specified standardized candidate dose-response models, as described in Table 13 will be utilized to examine the dose-response relationship. For each endpoint separately, a statistically significant dose response relationship will be declared if at least one model is identified by the MCP-Mod method to be statistically significant at the level $\alpha = 0.05$ one-sided. The fitted dose-response curves will be presented graphically for all statistically significant models along with confidence bands. The MED will be identified for each statistically significant model based on the pre-specified clinical meaningful target of a 20% target difference from placebo. The weighted MED across all significant models will be calculated, with weights being the inverse of model AIC.

Table 13. Candidate Models

Model	$f(d,\theta)$ d = dose, ϑ = Model Parameters	$f^0(d,\theta)$ Standardized Model	Initial Value(s) for Parameter(s)
Linear	$E_0 + \delta d$	d	NA
Exponential	$E_0 + E_1\left[exp\left(\frac{d}{\delta}\right) - 1\right]$	$exp\left(\frac{d}{\delta}\right) - 1$	δ = 10
Logistic	$E_0 + \frac{E_{max}}{1 + exp\left(\frac{ED_{50} - d}{\delta}\right)}$	$\frac{1}{1 + exp\left(\frac{ED_{50} - d}{\delta}\right)}$	ED_{50} = 15, δ = 5
EMax	$E_0 + \frac{E_{max} d}{ED_{50} + d}$	$\frac{d}{ED_{50} + d}$	ED_{50} = 15
sigEMax	$E_0 + \frac{E_{max} d^h}{ED_{50}{}^h + d^h}$	$\frac{d^h}{ED_{50}{}^h + d^h}$	ED_{50} = 15, h = 2.5
Quadratic	$E_0 + \beta_1 d + \beta_2 d^2$	$d + \frac{\beta_2}{\lvert\beta_1\rvert} d^2$	δ = -0.02

10.1.5 Missing Data Imputation

Missing data will be imputed using the following methods for the efficacy analyses:

- Non-Responder Imputation (NRI): the NRI analysis will categorize any subject who does not have evaluation during a specific visit window as a non-responder for that visit. The exception is when the subject is a responder both before and after a specific visit window, in which case the subject will be categorized as a responder for the visit. Only observations within the same analysis period will be used. The NRI will be the primary approach in the analyses of categorical variables.

- Last Observation Carried Forward (LOCF): The LOCF analyses will carry forward the most recent previous non-missing value (last completed non-missing value, for composite endpoint) to impute missing data at later visits in the same period. Baseline efficacy evaluations will not be carried forward. LOCF will be the primary approach in the analyses of continuous variables, and the secondary approach in the analysis of categorical variables.
- As-Observed Cases (OC): The as-observed analysis will not impute values for missing evaluations, and thus a subject who does not have an evaluation on a scheduled visit will be excluded from the as-observed analysis for that visit. As-observed analysis will be the secondary approach in the analysis of continuous variables.
- Multiple Imputation (MI): The MI approach will be used for sensitivity analyses on the Week 16 endpoints of EASI percent change from Baseline, EASI 75, and the Investigator Global Assessment (IGA) response of 0 (clear) or 1 (almost clear). If MI is not applicable due to the nature of the data (e.g., MCMC algorithm does not converge), logistic regression or mixed effect model repeat measurement (MMRM) methods will be applied as the sensitivity approach, whichever applicable.

The Multiple Imputation analysis will be carried out in three steps.

- Imputation of missing data. The imputation will be generated for each efficacy endpoint measurement. The variables to be included in the imputation model are: Baseline EASI (or IGA), treatment group, and measurements at each visit up to the end of the analysis period. For each endpoint, 20 'complete' datasets will be generated using SAS PROC MI. The imputed post-baseline measurements will be rounded to the same precision as the observed data before the determination of responder status (e.g., EASI 75 and EASI percentage change from Baseline).
- Analysis of imputed data sets. Using the corresponding statistical methods for the categorical and continuous endpoints, the imputed endpoints from Step 2 will be analyzed using each of the 20 datasets.

- Synthesis of imputation and analysis results. SAS PROC MIANALYZE will be used to generate the final inferences of the risk difference between treatment groups.

Of note, subjects who take rescue upadacitinib 30 mg or take rescue topical treatment in Period 2 will have their efficacy assessments taken from the date after the first dose of rescue excluded from the analyses, as the result, the subjects will be considered as non-responders in the NRI approach, and have their last observations on or prior to the date of rescue carried forward, and be removed from the OC approach.

10.2 Analysis of Daily Efficacy Measurements

For daily efficacy CCI [redacted]

For Period 1 or the entry value for Period 2, the rolling weekly average score is calculated starting from the first day when a measurement was available to the last day with a measurement, Let P_{m-6}, P_{m-5},…, P_{m-1}, P_m be the daily total scores from day m – 6 to day m, and N_m be the number of days with non-missing pain scores from day m – 6 to day m, then the rolling weekly average for Day m is:

$$\frac{\sum_{i=m-6}^{m} P_i}{N_m}$$

If the values from four or more days of the seven-day period are missing, then the rolling weekly average of Day m will be set to missing. If more than one assessment is included on the same day, the assessment associated with the worst condition on that day will be chosen as the daily score.

Analysis value for a given visit will be selected from rolling averages based on analysis window conventions, after dropping the rolling averages from Day 1 – 6 for Actigraphy, and Day 1 – 7 for other assessments, in each period.

For the visits where the questionnaire is completed at the site, scores from single clinic visits will be used as the analysis value.

10.3 Primary Efficacy Analysis

The primary efficacy variable is the percent (%) change from Baseline in EASI score at Week 16. The analysis of the primary efficacy variable will be carried out in ITT_1 with the LOCF missing data imputation. The dose-response relationship will be evaluated by MCP-Mod method.

10.4 Secondary Efficacy Analyses

The secondary efficacy variables include:

- Proportion of subjects achieving an EASI 75 response, defined as at least a 75% reduction in EASI score, at Week 16 relative to Baseline
- Proportion of subjects achieving an IGA 0 or 1 response at Week 16
- Percent change from Baseline in pruritus numerical rating scale (NRS) by week
- Percent change in EASI score from Baseline at Week 8
- Percent change in SCORAD score from Baseline at Weeks 8 and 16
- Proportion of subjects achieving EASI 50/75/90 response at Weeks 8 and 16
- Proportion of subjects achieving SCORAD 50/75/90 response at Weeks 8 and 16
- Change from Baseline in BSA at Week 16
- Percent change from re-randomization in EASI score in Period 2
- Time to loss of EASI 50 response among those who were re-randomized as EASI 75 responders at Week 16

- Summary of EASI 75 at all visits in Period 2 among those who were re-randomized as EASI 75 non-responders at Week 16
- Proportion of subjects with reduction of ≥ 4 from Baseline in pruritus NRS among subjects with a baseline of ≥ 4

10.5 Exploratory Variables

The following endpoints will be analyzed at all visits or weeks that the assessments are collected, except for daily assessments where results based on a rolling average by week will be presented.

- Time to EASI 50/75/90 and IGA "0" or "1" response in Period 1.
- Proportion of subjects achieving EASI 50/75/90/100 response.
- Proportion of subjects achieving SCORAD 50/75/90 response.
- Change from Baseline in POEM at Weeks 4, 16, 24 and 40.
- Change from Baseline at Weeks 4, 16, 24 and 40 in MOS Sleep Scale (Sleep Disturbance Score, Snoring Score, Short of Breath Score, Sleep Adequacy Score, Sleep Somnolence Score, Sleep Problems Index 1 Score, Sleep Problems Index 2 Score, Optimal Sleep Score, and Sleep Quantity Score).
- Change from Baseline at Weeks 16 and 40 in Asthma Symptoms Questionnaire.
- Change from Baseline at Weeks 16 and 40 in Daytime Nasal Symptoms Questionnaire.
- Change from Baseline in total sleep time per night (TST min), Sleep Efficiency (%), Wake After Sleep Onset (WASO), number of scratching events per hour, mean activity during rest (sleep) periods as measured by actigraphy, by week until Week 40 and by visit thereafter.
- Change from Baseline in ADERM-SS and ADERM-IS by week until Week 40 and by visit thereafter.
- Change from Baseline in PGIS.
- Proportion of subjects with a DLQI 0 or 1 response.

- Change from Baseline in DLQI.
- Change from re-randomization in EASI score in Period 2.

Primary and secondary variables will also be evaluated at all scheduled visits.

The efficacy of rescue will be summarized for subjects who received the rescue medication.

Due to collection issues in DLQI Question 7, the corresponding secondary endpoints have been changed to exploratory variables and the answers to Question 7 will be removed from the analyses. Of note, correction to the DLQI ePRO device has been corrected while the study is ongoing. Additional analyses based on correct DLQI data may be performed.

The EASI overall score, IGA and pruritus NRS score will be summarized at Baseline and Week 16 for subjects from Japan, with the percentage changes in EASI and pruritus NRS from baseline to Week 16.

In addition to the overall EASI score, body region EASI scores will be analyzed. The analysis will include all the four body component scores for head/neck, trunk, upper and lower extremities. Body component EASI scores are determined by multiplying the total of erythema, edema/papulation, excoriation, and lichenification scores within each body region by the area score.

EASI body region score = (erythema score + edema/papulation score + excoriation score + lichenification score)*area score

In addition, the SCORAD affected areas of head/neck and genitals will be analyzed along with the overall score.

Information on the scoring and missing value handling the SCORAD and Patient Reported Outcomes (PROS) is presented in Appendix A and Appendix B. The types of PROS include POEM, MOS, pruritus NRS, PGIS, ADerm-SS, ADerm-IS, Asthma

Symptoms Questionnaire, Daytime Nasal Symptom Questionnaire and ActiGraphy measurements.

MCP-Mod will also be performed for proportion of subjects achieving EASI 75 and achieving IGA 0 or 1 at Week 16.

10.6 Handling of Multiplicity

Multiplicity adjustments are not performed in this Phase 2b study.

10.7 Efficacy Subgroup Analysis

To evaluate the impact of baseline characteristics, subgroup analyses will be performed at Week 16 for the primary efficacy variable, percent change in EASI from baseline, and two of the secondary efficacy variables, proportion of subjects achieving EASI 75 and proportion of subjects achieving IGA of 0 or 1. The subgroups are defined by each of the demographic factors on age, gender, body mass index (BMI), race, weight and geographical region as shown in Table 14 below. Additional subgroup analysis based on baseline characteristics may be conducted.

If any of the resulting subgroups has fewer than 5 subjects, the subgroup analyses for that variable will not be presented.

Table 14. Subgroup Analysis

Subgroup Factor	Categories
Age	< 40 years old
	≥ 40 – < 65 years
	≥ 65 years old
Sex	Male
	Female
BMI	< 25
	≥ 25 – < 30
	≥ 30 – < 40
	≥ 40
Race	White
	Asian
	Other
Weight	< median
	≥ median
Geographic region	US/PR/Canada
	EU/AUS
	Japan

11.0 Safety Analysis

11.1 General Considerations

Safety analyses will include adverse events, laboratory, and vital sign measurements. Safety summaries will be provided using the safety population in each period. Comparison between each upadacitinib group and placebo group will be performed for change from baseline laboratory and vital sign parameters in the Safety_1 Population, using one-way ANOVA.

Missing safety data will not be imputed.

11.2 Analysis of Adverse Events

Treatment-emergent adverse events (TEAEs) will be summarized in each period.

Pre-treatment adverse event and protocol-defined non SAE will be presented in a data listing. TEAEs will also be summarized separately for the time period when subjects receive upadacitinib monotherapy of each dose level across Period 1 and Period 2, the time period when subjects receive concomitant mometasone furoate, and the time period when subjects receive concomitant triamcinolone acetonide.

11.2.1 Treatment-Emergent Adverse Events

A treatment-emergent Adverse Event (TEAE) is defined as an adverse event with an onset or worsening date that is on or after the first dose of study drug, and no more than 30 days after the last dose of study drug.

AEs where the onset date is the same as the study drug start date are assumed to be treatment-emergent. If an incomplete onset date was collected for an adverse event, the event will be assumed to be treatment-emergent unless there is other evidence that confirms that the event was not treatment-emergent (e.g., the event end date was prior to the study drug start date).

TEAEs for the safety population in each period are defined as an event with a start date on or after the date of first study drug dose in each period and up to 30 days after the last dose of study drug in the analysis period, or prior to the date of first dose in the subsequent period for subjects who entered in to the subsequent period.

Adverse event data will be coded using the Medical Dictionary for Drug Regulatory Activities (MedDRA) version 20.0 or most up to date version.

11.2.1.1 Adverse Event Overview

The number and percentage of subjects experiencing TEAEs will be summarized for each treatment group for the following AE categories.

- Any TEAE
- Any treatment-emergent serious adverse events (TESAE)
- Any TEAE leading to discontinuations of study drug

- Any severe TEAE
- Any reasonably possibly related TEAE
- Any reasonably possibly related TESAE
- Any TEAE leading to death
- TEAEs of Special Interest
- Any Death

11.2.1.2 Adverse Events by System Organ Class and Preferred Term

TEAEs will be summarized and presented by system organ class (SOC) and preferred terms (PT). The SOCs will be presented in alphabetical order, and the PTs will be presented in alphabetical order within each SOC.

A subject who reports more than 1 AE in different SOCs will be counted only once in the overall total. A subject who reports 2 or more different PTs within the same SOC, will be counted only once in the SOC total. A subject who reports more than 1 AE with the same PT will be counted only once for that PT using the most extreme incidence.

11.2.1.3 Adverse Events by Maximum Severity

TEAEs will be summarized by maximum severity. If a subject has an AE with an unknown severity, then the subject will be counted in the severity category of unknown, even if the subject has another occurrence of the same event with a severity present. The only exception is that if the subject has another occurrence of the same AE with the most extreme severity (severe). In this case, the subject will be counted under the severe category.

11.2.1.4 Adverse Events by Maximum Relationship

TEAEs will be summarized by maximum relationship to study drug, as assessed by the investigator. If a subject has an AE with an unknown relationship, then the subject will be counted in the relationship category of "unknown," even if the subject has another occurrence of the same event with a relationship present. The only exception is if the

subject has another occurrence of the same AE with a relationship assessment of reasonable possibility in which case, the subject will be counted under reasonable possibility category.

11.2.1.5 Frequent (≥ 5%) Adverse Events and Reasonably Possibly Related Adverse Events by System Organ Class and Preferred Term

TEAEs and reasonably possibly related AEs occurring for more than 5% of the subjects in any treatment groups will be summarized by MedDRA PT in decreasing frequency separately.

11.2.1.6 Adverse Events of Special Interests

The AESI categories will be identified by the following search criteria per Standard MedDRA Queries (SMQs)/Company MedDRA Queries (CMQs) specified in Table 15.

Table 15. AESI for Upadacitinib with SMQs/CMQs Searches

AESI	Type of MedDRA Query	Broad or Narrow Search	SMQ/CMQ Search Criteria
Serious Infections	CMQ		"Infections" – Subset for SAEs
Opportunistic Infection	CMQ		"Opportunistic Infection"
Malignancy	SMQ	Narrow	"Malignancies"
Non-Melanoma Skin Cancer (NMSC)	SMQ	Broad	Skin Malignant tumours (Broad SMQ) removing Melanoma CMQ
Malignancy excluding NMSC			Malignancy Narrow SMQ and removing NMSC output
Lymphoma	SMQ		"Malignant Lymphomas"
Hepatic Disorder	SMQ	Narrow	"Drug Related Hepatic Disorders"
Gastrointestinal Perforations	SMQ	Narrow	"Gastrointestinal Perforation"
Anemia	CMQ		"Non-Hemolytic and Non-Aplastic Anemias"
Neutropenia	CMQ		"Hematological Toxicity – Neutropenia"
Lymphopenia	CMQ		"Hematological Toxicity – Lymphopenia (Veliparib Product Specific)"
Herpes Zoster	CMQ		"Herpes Zoster"
Rhabdomyolysis/myopathy	SMQ	Narrow	"Rhabdomyolysis/Myopathy"
Creatine Phosphokinase (CPK Elevation)	PT		Search only for the PT of "Blood creatine phosphokinase increased"
Renal Dysfunction	SMQ	Narrow	"Acute Renal ,Failure"
Tuberculosis	CMQ		"Tuberculosis"
Adjudicated cardiovascular events[a]	Output from CAC		
Cardiac Arrhythmias	SMQ	Narrow	"Torsade de pointes/QT Prolongation" "Tachyarrhythmias"

a. Reviewed and adjudicated by an independent Cardiovascular Adjudication Committee in a blinded manner.

Additional AEs may be considered for tabulation/summary based on recommendations from Clinical and Safety as deemed appropriate.

11.2.1.7 Adverse Events by 100 Patient Years

TEAEs will be summarized by event rate per 100 subject years, defined as

100 * (Number of TEAEs)/(Total Patient Years)

where total patient years is defined as the sum of the study drug exposure (defined in Section 9.1) of all subjects normalized by 365.25, and rounded to 1 decimal place. Event rate (per 100 patient years) for AEs of special interest will also be summarized for the combined safety analysis of Period 1 and Period 2.

11.2.2 Serious Adverse Events (Including Deaths) and Adverse Events Leading to Study Drug Discontinuation

All TESAEs, deaths, and TEAEs leading to discontinuation of study drug will be summarized. The number and percentage of subjects experiencing SAEs (including deaths) and adverse events leading to discontinuation of study drug will be tabulated by SOC and PT for each treatment group.

11.2.3 Listing of Adverse Events

The following additional summaries will be provided for treatment-emergent adverse events, unless otherwise specified.

- List of subject numbers associated with each PT for TEAEs.
- List of subject numbers associated with each PT for TEAEs assessed by the investigator as reasonable possibility of being related to study drug.
- Listing of pretreatment AEs.
- Listing of TESAEs.
- Listing of TEAEs that led to discontinuation of study drug.
- Listing of all deaths.
- Listing of treatment-emergent AESIs.

11.3 Analysis of Laboratory Data

For the assessments of laboratory data, values observed more than 30 days after the last dose of study drug will be excluded.

All laboratory parameters to be collected in this study are listed in Table 16 below. Laboratory parameters will be reported using the standard international (SI) units.

Table 16. List of Laboratory Variables

Hematology	Chemistry	Urinalysis
White Blood Cell (WBC) Count Red Blood Cell (RBC) Count Hemoglobin Hematocrit Platelets count Neutrophils Basophils Eosinophils Lymphocytes Monocytes Reticulocytes* Bands	Total Bilirubin Alkaline Phosphatase (ALP) Serum glutamic oxaloacetic transaminase (SGOT/AST) Serum glutamic pyruvic transaminase (SGPT/ALT) Total Protein Albumin Glucose Triglycerides Blood Urea Nitrogen (BUN) Creatinine Uric acid Sodium Potassium Calcium Inorganic Phosphorus Creatine Phosphokinase (CPK) Chloride Bicarbonate Cholesterol LDL cholesterol HDL cholesterol	Specific Gravity pH Protein* Glucose* Ketones* Blood* Nitrite* Microscopic Examination*
		Other
		hsCRP HBs Ab* HBc Ab* Hbs Ag* HBV DNA PCR reflex only* HCV Ab* HCV RNA reflex only* QuantiFERON-TB Gold*

* The parameters will only be presented in data listing.

11.3.1 Variables and Criteria Defining Abnormality

Criteria for Potentially Clinically Significant chemistry and hematology variables are listed in Table 17 below.

Table 17. Definitions of Potentially Clinically Significant Chemistry and Hematology Variables Based on CTCAE Grades

Test	Grade 1	Grade 2	Grade 3	Grade 4
SGPT/ALT	> ULN – 3.0 × ULN	> 3.0 – 5.0 × ULN	> 5.0 – 20.0 × ULN	> 20.0 × ULN
SGOT/AST	> ULN – 3.0 × ULN	> 3.0 – 5.0 × ULN	> 5.0 – 20.0 × ULN	> 20.0 × ULN
ALP	> ULN – 2.5 × ULN	> 2.5 – 5.0 × ULN	> 5.0 – 20.0 × ULN	> 20.0 × ULN
TBL	> ULN – 1.5 × ULN	> 1.5 – 3.0 × ULN	> 3.0 – 10.0 × ULN	> 10.0 × ULN
Creatinine increased	> ULN – 1.5 × ULN	> 1.5 – 3.0 × ULN	> 3.0 – 6.0 × ULN	> 6.0 × ULN
CPK increased	> ULN – 2.5 × ULN	> 2.5 – 5.0 × ULN	> 5.0 – 10.0 × ULN	> 10.0 × ULN
Hemoglobin decreased	< LLN – 100.0 g/L	< 100.0 – 80.0 g/L	< 80.0 g/L	N/A
Neutrophil count decreased	< LLN – 1.5 × 10^9/L	< 1.5 – 1.0 × 10^9/L	< 1.0 – 0.5 × 10^9/L	< 0.5 × 10^9/L
WBC decreased	< LLN – 3.0 × 10^9/L	< 3.0 – 2.0 × 10^9/L	< 2.0 – 1.0 × 10^9/L	< 1.0 × 10^9/L
Lymphocyte count decreased	< LLN – 0.8 × 10^9/L	< 0.8 – 0.5 × 10^9/L	< 0.5 – 0.2 × 10^9/L	< 0.2 × 10^9/L

Definitions of Potential Hy's Law Cases

According to FDA's Guidance for Industry "Drug-Induced Liver Injury: Premarketing clinical evaluation (July 2009), when aminotransferase (AT) abnormalities indicating hepatocellular injury are accompanied by evidence of impaired hepatic function (bilirubin elevation > 2 × ULN), in the absence of evidence of biliary obstruction (i.e., significant elevation of ALP) or some other explanation of the injury (e.g., viral hepatitis, alcohol hepatitis), the combined finding (i.e., Hy's Law cases) represents a signal of a potential for the drug to cause severe DILI.

11.3.2 Statistical Methods

11.3.2.1 Analysis for Continuous Laboratory Data

Mean changes from baseline in all continuous laboratory parameters in each visit in Period 1 will be summarized with the number of observations, mean, standard deviation and median. The treatment comparison between groups will be using a one-way ANOVA. Percent change from baseline for selected lab parameters will also be summarized. Summaries for the baseline and final value will be presented for subjects

who have both baseline and post-baseline values. Categorical data will be summarized using frequencies and percentages.

11.3.2.2 Shift Table Analyses

Baseline and post-baseline laboratory observations will be categorized as Grade 1, Grade 2, Grade 3, and Grade 4 according to National Cancer Institute Common Toxicity Criteria (NCI CTC) AE version 4.03 grades (publish date: June 14, 2010) (See definition in Table 17). Shift tables for changes from Baseline according to the NCI CTC toxicity grade will be provided for each hematology, clinical chemistry and urinalysis parameter.

11.3.2.3 Potentially Clinically Significant Laboratory Values

Criteria for identify Potentially Clinically Significant laboratory values are listed in the table above. Of note, a post-baseline value has to be more extreme than baseline value to be considered.

Frequencies and percentages of subjects with post Baseline lab values that are Grade 3 or above according to the CTC toxicity criteria, will be summarized. A listing of the subjects with laboratory values that are CTC toxicity Grade 3 or above will be provided. For each of these subjects, the whole course of the respective parameter will be listed.

11.3.2.4 Liver Function Tests

Additional summaries will be presented for liver function tests including ALT or SGPT, AST or SGOT, alkaline phosphatase, and total bilirubin. Each laboratory value will be categorized as follows:

- $< 1.5 \times$ ULN
- $\geq 1.5 \times$ ULN – $< 3.0 \times$ ULN
- $\geq 3.0 \times$ ULN – $< 5.0 \times$ ULN
- $\geq 5.0 \times$ ULN – $< 10.0 \times$ ULN
- $\geq 10.0 \times$ ULN – $< 20.0 \times$ ULN
- $\geq 20.0 \times$ ULN

For the purpose of assessing for potential Hy's law cases, the frequencies and percentages of subjects with post baseline liver specific function test values that meet the following criteria of potential clinical interest should be presented:

- ALT ≥ 3 × ULN
- ALT ≥ 5 × ULN
- ALT ≥ 10 × ULN
- ALT ≥ 20 × ULN
- AST ≥ 3 × ULN
- AST ≥ 5 × ULN
- AST ≥ 10 × ULN
- AST ≥ 20 × ULN
- TBL ≥ 2 × ULN
- Alkaline phosphatase ≥ 1.5 × ULN
- ALT and/or AST ≥ 3 × ULN and concurrent TBL ≥ 1.5 × ULN
- ALT and/or AST ≥ 3 × ULN and concurrent TBL ≥ 2 × ULN

11.4 Analysis of Vital Signs and Weight

11.4.1 Variables and Criteria Defining Abnormality

Definition of Potentially Clinically Significant vital signs variables based on Criteria are listed in Table 18 below.

Table 18. Definition of Potentially Clinically Significant Vital Signs Variables

Vital Sign	Category	Criteria for Potential Clinically Significant Vital Signs
Systolic blood pressure	Low	Value ≤ 90 mmHg and decrease ≥ 20 mmHg from Baseline
	High	Value ≥ 160 mmHg and increase ≥ 20 mmHg from Baseline
Diastolic blood pressure	Low	Value ≤ 50 mmHg and decrease ≥ 15 mmHg from Baseline
	High	Value ≥ 105 mmHg and increase ≥ 15 mmHg from Baseline
Pulse	Low	Value ≤ 50 bpm and decrease ≥ 15 bpm from Baseline
	High	Value ≥ 120 bpm and increase ≥ 15 bpm from Baseline
Respiratory Rate	Low	< 10 rpm
	High	> 24 rpm
Body temperature	High	> 39.0°C (102.3°F)
Weight	High	> 7% increase from baseline
	Low	> 7% decrease from baseline

11.4.2 Statistical Methods

Changes from Baseline to post-baseline visits will be summarized with the baseline mean, the visit mean, change from baseline mean, standard deviation, and median. An ANOVA model with only treatment as a factor, not controlling for baseline, will be used to test statistical significance for the change from baseline mean, and present confidence intervals for the difference between each of the upadacitinib treatment group and placebo in Period 1. The number and percentage of subjects meeting the criteria for potentially clinically significant vital sign values will be summarized.

Vital sign results satisfying the criteria for potentially clinically significant vital sign findings will be identified in a listing. Of note, a post-baseline value has to be more extreme than baseline value to be considered.

11.5 Analysis of ECG Parameters

ECG is collected at baseline visit, and Week 16. ECG findings will be summarized by treatment group for each parameter and visit. Summaries will include n (%) of patients in following categories:

- Normal
- Abnormal – Not Clinically Significant
- Abnormal – Clinically Significant
- Unable To Evaluate
- Missing

12.0 Summary of Changes

12.1 Summary of Changes Between the Latest Version of Protocol and the Current SAP

The protocol planned statistical tests for baseline variables were removed sine it is not meaningful to conduct hypothesis test of randomization.

In the protocol, five models were proposed to be considered for MCP-Mod: linear, E_{max}, exponential, logistic and sigE_{max}. In the SAP, quadratic model was added as an additional candidate model.

In the protocol, DLQI is a secondary endpoint. Due to collection issues in DLQI Question 7, the corresponding secondary endpoints have been changed to exploratory variables and the answers to Question 7 will be removed from the analyses. Of note, correction to the DLQI ePRO device has been implemented while the study is ongoing. Additional analyses based on correct DLQI data may be performed.

Proportion of subjects achieving an improvement of 4 or more points from Baseline in pruritus NRS among subjects with a baseline of at least 4 was added as a secondary endpoint.

12.2 Summary of Changes Between the Previous Version and the Current Version of the SAP

Not applicable. This is the first version of SAP.

13.0 References

1. Greenland S, Robins M. Estimation of a common effect parameter from sparse follow-up data. Biometrics. 1985;41(1):55-68.
2. Pinheiro J, Bornkamp B, Bretz F. Design and analysis of dose-finding studies combining multiple comparisons and modeling procedures. J Biopharm Stat. 2006;16(5):639-56.
3. Bretz F, Pinheiro J, Branson M. Combining multiple comparisons and modeling techniques in dose-response studies. Biometrics. 2005;61(3):738-48.

Appendix A. SCORAD Scoring Algorithm

Appendix A. SCORAD Scoring Algorithm

Table of Contents

0.0 Questionnaire

Body Area Affected:

The extent of AD is assessed as a percentage of each defined body area. To help in determining this extent, the sites affected by eczema are shaded on a drawing of a body. The rule of 9 is used to calculate the affected area (A) as a percentage of the whole body:

- Head and neck 9%
- Upper limbs 9% each
- Lower limbs 18% each
- Anterior trunk 18%
- Back 18%
- Genitals 1%

The score for each area is added up. The total area is 'A,' which has a possible maximum of 100%.

Symptom Severity:

A representative area of eczema is selected. In this area, the intensity of each of the following 6 specific symptoms is assessed as none (0), mild (1), moderate (2) or severe (3).

- Redness
- Swelling
- Oozing/crusting
- Scratch marks

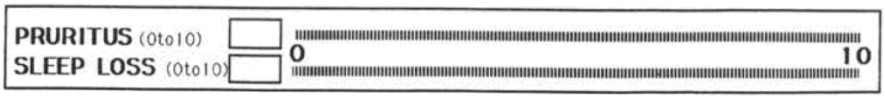
- Skin thickening (lichenification)
- Dryness (this is assessed in an area where there is no inflammation)

The scores for these 6 specific symptoms should be added, for a maximum of 18 total points, assigned as "B" in the overall SCORAD calculation.

Subjective Symptoms:

Subjective assessment of itch and sleeplessness is recorded for each symptom by the subject on a visual analogue scale (VAS), where 0 is no itch (or sleeplessness) and 10 is the worst imaginable itch (or sleeplessness), with a maximum possible score of 20. This parameter is assigned as "C" in the overall SCORAD calculation.

Visual analog scale (average for the last 3 days or nights)



(10 cm in length)

1.0 Scoring Algorithm

The SCORAD is calculated as: A/5 + 7B/2 + C.

A-Total affected area
B-Symptom severity
C- Subjective assessment of itch and sleeplessness.

2.0 Missing Value Handling

If any of the components is missing, then the overall score will be treated as missing.

3.0 Efficacy Variables

- Changes from Baseline for the overall SCORAD score and the affected area scores of head/neck and genitals

- Achievement on pre-specified percentages of improvement, SCORAD 50/75/90 .

4.0 Reference

Dermatology in 1993; 186(1): 23-31 Consensus Report of the European Task Force on Atopic Dermatitis.

Appendix B. Patient Reported Outcome Scoring

Appendix B. Patient Reported Outcome Scoring

Table of Contents

1.0 Patient Oriented Eczema Measure (POEM)

1.1 Questionnaire

POEM
Patient-Oriented Eczema Measure

The University of Nottingham
UNITED KINGDOM · CHINA · MALAYSIA

POEM for self-completion

Patient Details: _____

_____

_____

_____ Date: _____

Please circle one response for each of the seven questions below about your eczema. Please leave blank any questions you feel unable to answer.

1. Over the last week, on how many days has your skin been itchy because of your eczema?

No days | 1-2 days | 3-4 days | 5-6 days | Every day

2. Over the last week, on how many nights has your sleep been disturbed because of your eczema?

No days | 1-2 days | 3-4 days | 5-6 days | Every day

3. Over the last week, on how many days has your skin been bleeding because of your eczema?

No days | 1-2 days | 3-4 days | 5-6 days | Every day

4. Over the last week, on how many days has your skin been weeping or oozing clear fluid because of your eczema?

No days | 1-2 days | 3-4 days | 5-6 days | Every day

5. Over the last week, on how many days has your skin been cracked because of your eczema?

No days | 1-2 days | 3-4 days | 5-6 days | Every day

6. Over the last week, on how many days has your skin been flaking off because of your eczema?

No days | 1-2 days | 3-4 days | 5-6 days | Every day

7. Over the last week, on how many days has your skin felt dry or rough because of your eczema?

No days | 1-2 days | 3-4 days | 5-6 days | Every day

1.2 Scoring Algorithm

Each of the seven questions carries equal weight and the responses are scored from 0 to 4 as detailed below:

- 0 = no days
- 1 = 1-2 days
- 2 = 3-4 days
- 3 = 5-6 days
- 4 = Every day

Then take the sum of seven responses.

1.3 Missing Value Handling

If one question is left unanswered or scored “unable to answer” this is scored 0 and the scores are summed and expressed as usual out of a maximum of 28. If more than one question is left unanswered or scored “unable to answer” the questionnaire is not scored. If two or more response options are selected for a single question, the response option with the highest score should be recorded.

1.4 Efficacy Variables

- The sum of the seven questions will be analyzed by change from Baseline.

1.5 References

1. The Patient-Oriented Eczema Measure: development and initial validation of a new tool for measuring atopic eczema severity from the patients' perspective. Charman C, Venn A, Williams HC. Arch Dermatol. 2004;140:1513-1519

2. Translating Patient-Oriented Eczema Measure (POEM) scores into clinical practice by suggesting severity strata derived using anchor-based methods. Charman CR, Venn AJ, Ravenscroft JC, Williams HC. Br J Dermatol. 2013; 169(6): 1326–1332.

3. The Patient-Oriented Eczema Measure in young children: responsiveness and minimal clinically important difference. Gaunt DM, Metcalfe C, Ridd M. Allergy 2016

4. EASI, (objective) SCORAD and POEM for atopic eczema: responsiveness and minimal clinically important difference. Schram ME, Spuls PI, Leeflang MM, Lindeboom R, Bos JD, Schmitt J. Allergy 2012 Jan; 67(1):99-106.

5. POEM a core instrument to measure symptoms in clinical trials: a HOME statement.Spuls PI, Gerbens LA, Simpson E, Apfelbacher CJ, Chalmers JR, Thomas KS, Prinsen CA, von Kobyletzki LB, Singh JA, Williams HC, Schmitt J; HOME initiative collaborators .Br J Dermatol. 2016.

2.0 Medical Outcomes Study (MOS)

2.1 Questionnaire

Your Sleep

For each of the following questions, please mark an ☒ in the one box that best describes your answer.

1. How long did it usually take for you to <u>fall asleep</u> during the <u>past 4 weeks</u>?

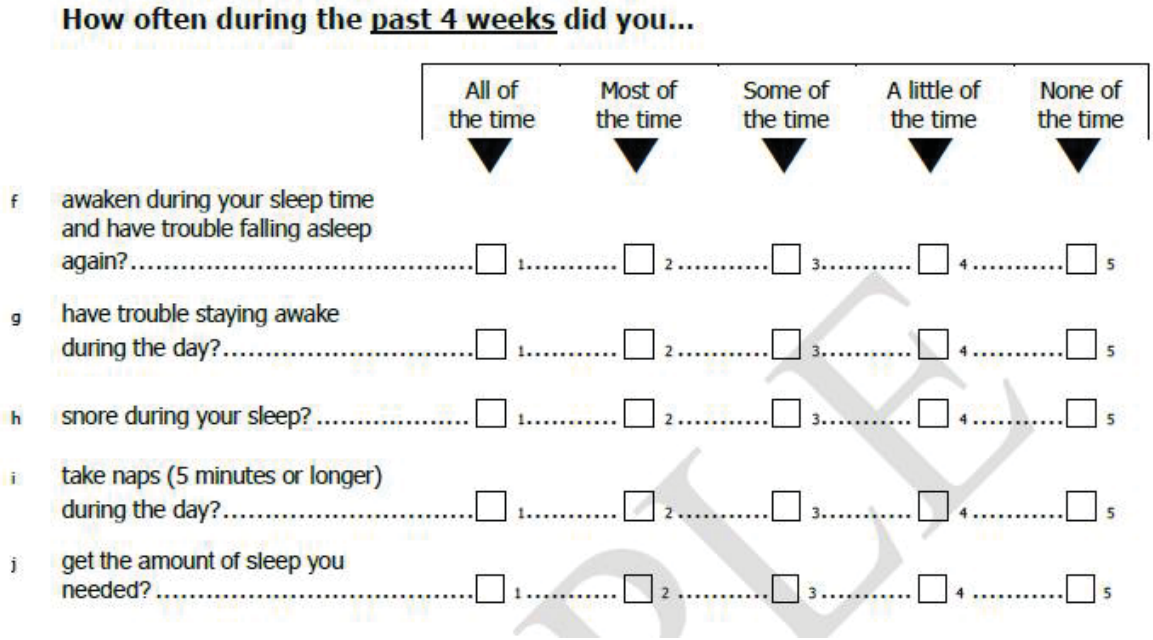
0-15 minutes	16-30 minutes	31-45 minutes	46-60 minutes	More than 60 minutes
☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

2. On the average, how many hours did you sleep <u>each night</u> during the <u>past 4 weeks</u>?

Write in number of hours per night: ☐☐

3. How often during the <u>past 4 weeks</u> did you...

	All of the time	Most of the time	Some of the time	A little of the time	None of the time
a	feel that your sleep was not quiet (moving restlessly, feeling tense, speaking, etc., while sleeping)?	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
b	get enough sleep to feel rested upon waking in the morning?	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
c	awaken short of breath or with a headache?	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
d	feel drowsy or sleepy during the day?	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
e	have trouble falling asleep?	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5



2.2 Scoring Algorithm

The 12-item MOS Sleep yields scores on six subscales, each consisting of one to four items

The subscales are:

- *Sleep Disturbance*. This four-item subscale addresses sleep initiation problems, as reported via the usual length of time to fall asleep and the frequency of trouble falling asleep, and sleep maintenance problems, as indicated by the frequency of experiencing sleep that was not "quiet" and the frequency of being awakened with trouble falling back to sleep.
    - Q#, 1, 3e, 3a, 3f

- *Sleep Quantity*. This single-item subscale represents the average number of hours the respondent sleeps each night.
    - Q: 2

- *Sleep Adequacy*. This two-item subscale asks about the frequency of awakening fresh and rested in the morning and the frequency of getting the amount of sleep needed.
    - Q: 3b

- *Somnolence*. Previously referred to as "Sleep Somnolence," this three-item subscale measures how frequently during the day the respondent feels drowsy, takes naps, and has trouble staying awake.
    - Q: 3d, 3i, 3g

- *Shortness of Breath or Headache*. This single-item subscale asks about the frequency of awakening during the night with shortness of breath or with a headache.Q: 3c

- *Snoring*. This single-item subscale addresses the frequency of snoring during sleep.
    - Q: 3h

The sub-scores and overall score are calculated by the external vender OPTUM, and thus no missing value algorithm is available in this document.

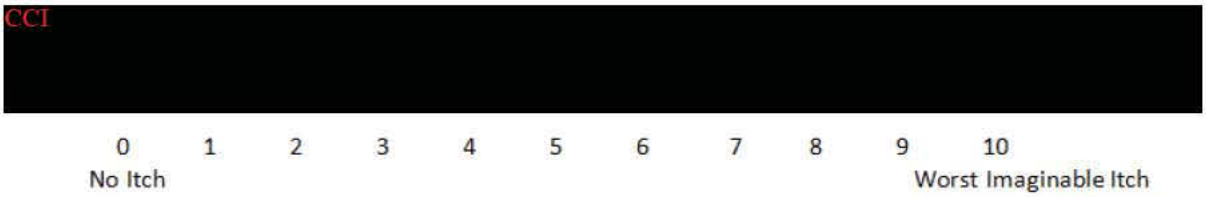
2.3 Efficacy Variables

- The six sub-scores and the overall score will be analyzed by change from Baseline.

3.0 Pruritus Numerical Rating Scale (NRS)

3.1 Questionnaire

Pruritus/Itch Numeric Rating Scale



3.2 Scoring Algorithm

The daily pruritus NRS will be collected from patients electronically every evening via a hand-held device provided to the subject at Screening. A rolling weekly average is calculated..

3.3 Missing Value Handling

Missing values will not be imputed for the daily pruritus NRS.

3.4 Efficacy Variable

- Rolling average will be analyzed as percent change from baseline.
- Proportion of subjects achievement of a 4-point reduction will be analyzed.

4.0 Patient Global Impression of Severity (PGIS)

4.1 CCI

CCI

4.2 Scoring Algorithm

PGIS is collected at site visits.

4.3 Missing Value Handling

Missing values will not be imputed for PGIS.

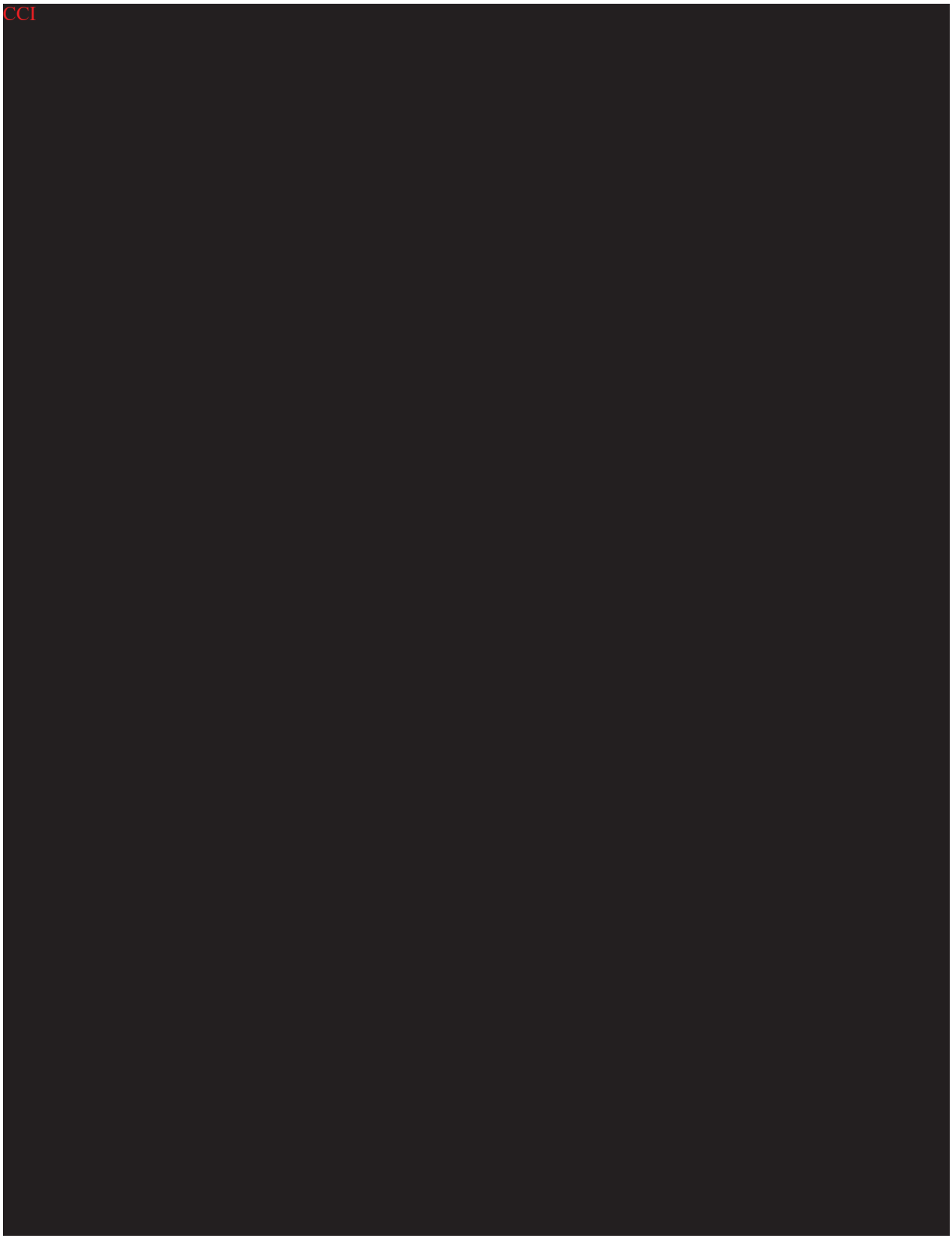
4.4 Efficacy Variable

PGIS will be analyzed by change from Baseline.

5.0 Atopic Dermatitis Symptom Scale (ADerm-SS)

5.1 CCI

CCI


CCI

5.2 Scoring Algorithm

CCI [redacted]

5.3 Missing Value Handling

Missing values will not be imputed for ADerm-SS. If any of Items CCI is missing then the CCI [redacted] Score will be treated as missing.

5.4 Efficacy Variable

- For each of Items CCI changes from Baseline will be analyzed using the rolling CCI [redacted] average.
- The CCI [redacted] Score will be analyzed by change from Baseline.

6.0 Atopic Dermatitis Impact Scale (ADerm-IS)

6.1 CCI

CCI

CCI

6.2 Scoring Algorithm

CCI

6.3 Missing Value Handling

Missing values will not be imputed for ADerm-IS. If any of Items CCI is missing then the CCI Score- will be treated as missing.

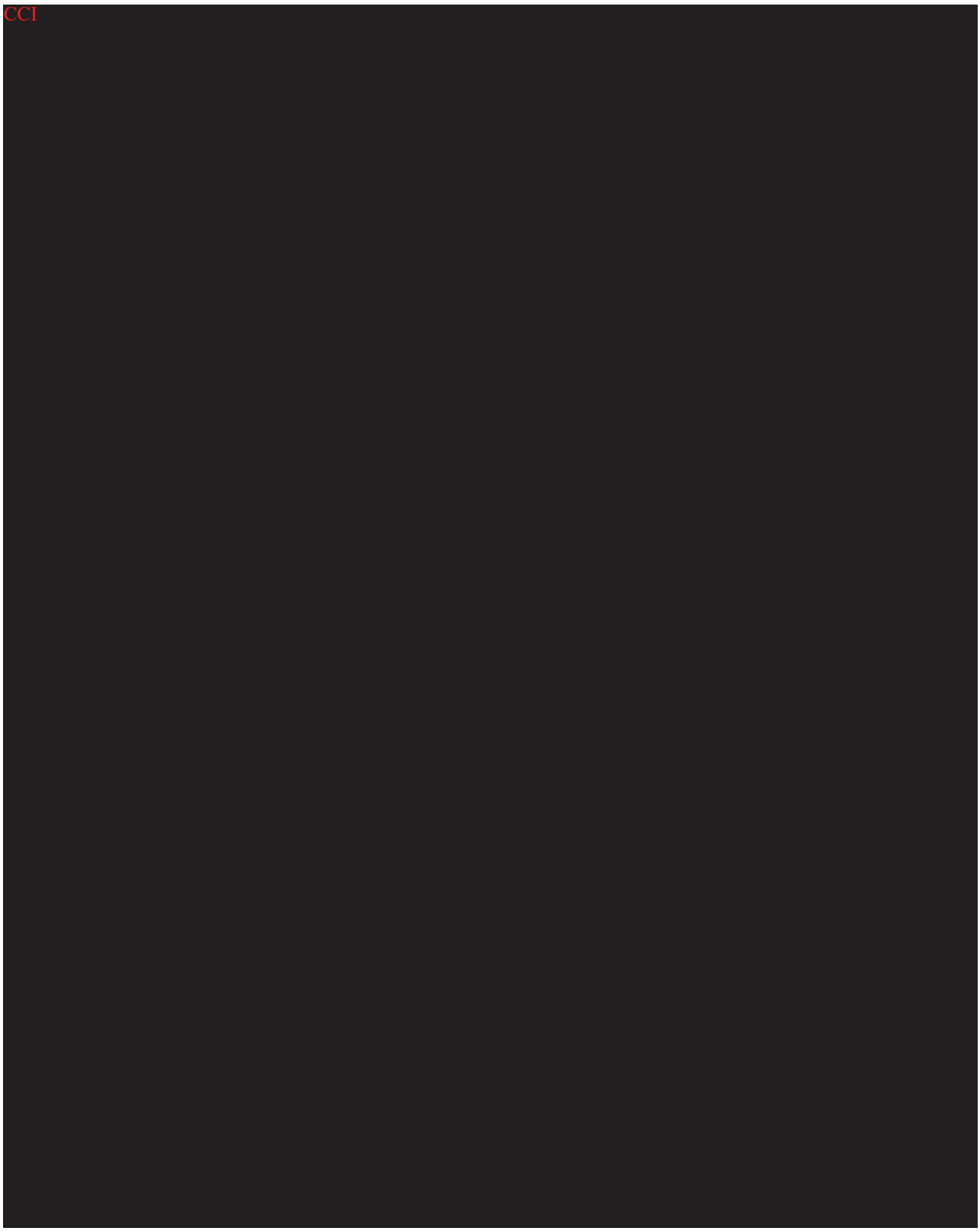
6.4 Efficacy Variable

- For each of Items CCI, changes from Baseline will be analyzed using the rolling CCI average.
- The CCI Score is analyzed by change from Baseline.

7.0 Asthma Symptoms Questionnaire

7.1 CCI

CCI



7.2 Scoring Algorithm

CCI [redacted]

7.3 Missing Value Handling

The missing values of individual items will not be imputed. If response CCI [redacted] CCI [redacted] is selected CCI [redacted], remaining items automatically become void. CCI [redacted] No missing value imputation is performed for individual Items CCI [redacted]

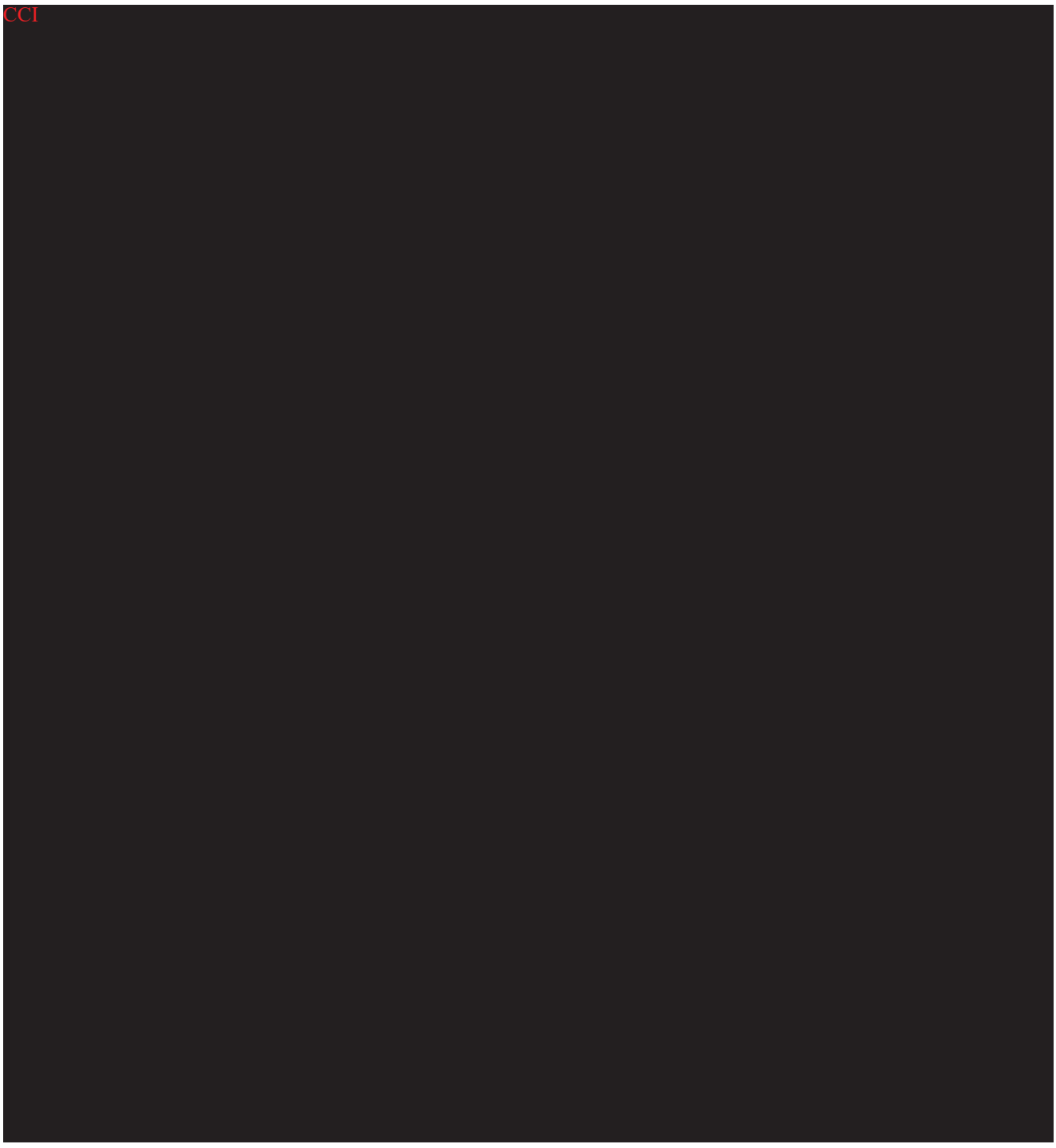
7.4 Efficacy Variables

- Items CCI [redacted] will be analyzed by frequency and percentage of response categories.
- Item CCI [redacted] and the summary score CCI [redacted] will be analyzed by change from Baseline. The analysis will be performed only for subjects who had a response CCI [redacted] at baseline.

8.0 Daytime Nasal Symptom Questionnaire

8.1 CCI

CCI



8.2 Scoring Algorithm

CCI [redacted]

8.3 Missing Value Handling

If any of the responses is missing then corresponding summary score will be treated as missing.

8.4 Efficacy Variables

- CCI [redacted] will be analyzed by frequency and percentage of response categories.
- CCI [redacted] Symptoms and CCI [redacted] Scores will be analyzed by change from Baseline. CCI [redacted]

9.0 ActiGraphy Measurements

Patient data on sleep activity and scratching events are collected daily with an electronic actigraphy device called Geneactiv Scratching Device. The measurements of interest downloaded from the device are total sleep time per night (TST min, in minutes), Sleep Efficiency (%), Wake After Sleep Onset (WASO, in minutes), and mean activity during rest (sleep), The parameter of interest related to scratching is events per hour. These five parameters will be analyzed by change from Baseline using the rolling weekly average.

10.0 Dermatology Life Quality Index (DLQI)

10.1 Questionnaire

DLQI is a simple 10-item questionnaire used to assess the handicap experienced by patients with various chronic skin diseases.

DERMATOLOGY LIFE QUALITY INDEX **DLQI**

Hospital No: Date: Score:

Name: Diagnosis:

Address:

The aim of this questionnaire is to measure how much your skin problem has affected your life OVER THE LAST WEEK. Please tick one box for each question.				
1.	Over the last week, how **itchy**, **sore**, **painful** or **stinging** has your skin been?	Very much A lot A little Not at all	☐ ☐ ☐ ☐	
1.	Over the last week, how **embarrassed** or **self conscious** have you been because of your skin?	Very much A lot A little Not at all	☐ ☐ ☐ ☐	
2.	Over the last week, how much has your skin interfered with you going **shopping** or looking after your **home** or **garden?**	Very much A lot A little Not at all	☐ ☐ ☐ ☐	Not relevant ☐
3.	Over the last week, how much has your skin influenced the **clothes** you wear?	Very much A lot A little Not at all	☐ ☐ ☐ ☐	Not relevant ☐
4.	Over the last week, how much has your skin affected any **social** or **leisure** activities?	Very much A lot A little Not at all	☐ ☐ ☐ ☐	Not relevant ☐
6.	Over the last week, how much has your skin made it difficult for you to do any **sport**?	Very much A lot A little Not at all	☐ ☐ ☐ ☐	Not relevant ☐
7a	Over the last week, has your skin prevented you from **working** or **studying**?	yes no	☐ ☐	Not relevant ☐

7 b	If "No", over the last week how much has your skin been a problem at **work** or **studying**?	A lot A little Not at all	☐ ☐ ☐	
8 .	Over the last week, how much has your skin created problems with your **partner** or any of your **close friends** or **relatives**?	Very much A lot A little Not at all	☐ ☐ ☐ ☐	Not relevant ☐
9 .	Over the last week, how much has your skin caused any **sexual difficulties**?	Very much A lot A little Not at all	☐ ☐ ☐ ☐	Not relevant ☐
1 0 .	Over the last week, how much of a problem has the **treatment** for your skin been, for example by making your home messy, or by taking up time?	Very much A lot A little Not at all	☐ ☐ ☐ ☐	Not relevant ☐

10.2 Scoring Algorithm

The DLQI measures six aspects of impaired dermatologic quality of life (Symptoms and feelings, Daily activities, Leisure, Work and School, Personal relationships, Treatment) and scores range from 0-30, with higher scores indicating more impaired quality of life.

Table 1. Scoring of DLQI questions

The scoring of each question is as follows:	
Very much	scored 3
A lot	scored 2
A little	scored 1
Not at all	scored 0
Not relevant	scored 0
Question unanswered	scored 0
Question 7: "prevented work or studying"	scored 3

The DLQI is calculated by summing the score of the 10 questions resulting in a maximum of 30 and a minimum of 0. The higher the score, the more quality of life is impaired. The DLQI can also be expressed as a percentage of the maximum possible score of 30

10.3 Missing Value Handling

For DLQI, missing values in individual questions are handled as follows.

- One Question unanswered this is scored 0 and the scores are summed and expressed as usual out of a maximum of 30.
- Two or more questions left unanswered – do not score questionnaire.
- If two or more response options (adjacent or non-adjacent) are checked for any question, the response option with the highest score will be entered into the database.
- If one item is missing from a two-item subscale that subscale should not be scored.
- If there is a response between two tick boxes, the lower of the two score options should be recorded. If there are responses between more than two tick boxes, the lowest of all the score options should be recorded.
- Question 7 is always counted as one question (for Abbott data management purposes only, the responses are coded in two sub-questions 7A and 7B.
    - When question 7A is "yes" then the score = 3 and question 7B is ignored.
    - When question 7A is "no":
      the score=2 if question 7B is "a lot"

      the score=1 if question 7B is "a little"

      the score=0 if 7B is "not at all"

      the score=missing if 7B is not checked

    - When question 7A is "not relevant" and 7B is not checked, the score=0.

    If 7B is checked:

      the score=2 if question 7B is "a lot"

      the score=1 if question 7B is "a little"

      the score=0 if question 7B is "not at all"

    - When question 7A is missing:
      the score=2 if question 7B is "a lot".

      the score=1 if question 7B is "a little".

      the score=0 if question 7B is "not at all" or "not relevant".

      the score=missing if 7B is not scored.

    - When question 7A is "no" and "not relevant", enter "not relevant" for question 7A.

10.4 Efficacy Variables

- Proportion of subjects with a DLQI 0 or 1 response.
- Change from Baseline in total DLQI score.

10.5 Adjustment Accommodating Device Error in this Study

During Period 1 of the study, there was an error in the device that Question 7.b) was displayed to a subject only when the subject answered "Yes" to Question 7.c). The error was corrected in August, with the baseline measurements being affected for all subjects, and the Period 1 post-baseline measurements being impacted for almost all subjects. In order to mitigate the bias caused by the device error, Items 7. a) and b) are excluded in calculating the total DLQI score for the Period 1 visits of all subjects.

10.6 References:

1. Wales College of Medicine, Department of Dermatology. Website: http://www.dermatology.org.uk/index.asp. Date accessed: May 20, 2008

2. Finlay A Y, Khan G K. Dermatology Life Quality Index(DLQI): A simple practical measure for routine clinical use. Clinical and Experimental Dermatology.1994;19:210-216

3. Shikiar R, Willian MK, Okun MM, Thompson CS, Revicki DA. The Validity and Responsiveness of Three Quality of Life Measures in the Assessment of Psoriasis Patients: Results of a Phase II Study. Health Qual Life Outcomes. 2006;4:71(27 September 2006). Open access available at www.hqlo.com.

4. Shikiar R, Bresnahan BW, Stone SP, Thompson C, Koo J, Revicki DA. Validity and reliability of patient reported outcomes used in Psoriasis: results from two randomized clinical trials. Health Qual Life Outcomes 2003, 1:53.

5. Willian MK, Chmiel J, Chen N. Dermatology Life Quality Index (DLQI) Scoring Convention (Abbott Internal Document). February 16, 2007.