Statistical Analysis Plan I4V-MC-JAIY Version 2

A Multicenter, Randomized, Doubles-Blind, Placebo-Controlled, Phase 3 Study to Evaluate the Efficacy and Safety of Baricitinib in Combination with Topical Corticosteroids in Adult Patients with Moderate to Severe Atopic Dermatitis

NCT03733301

Approval Date: 12-Aug-2019

1.   Statistical Analysis Plan for US and Japan: I4V-MC-JAIY (a): A Multicenter, Randomized, Double-Blind, Placebo-Controlled, Phase 3 Study to Evaluate the Efficacy and Safety of Baricitinib in Combination with Topical Corticosteroids in Adult Patients with Moderate to Severe Atopic Dermatitis

BREEZE-AD7

Confidential Information

The information contained in this document is confidential and the information contained within it may not be reproduced or otherwise disseminated without the approval of Eli Lilly and Company or its subsidiaries.

Note to Regulatory Authorities: This document may contain protected personal data and/or commercially confidential information exempt from public disclosure. Eli Lilly and Company requests consultation regarding release/redaction prior to any public release. In the United States, this document is subject to Freedom of Information Act (FOIA) Exemption 4 and may not be reproduced or otherwise disseminated without the written approval of Eli Lilly and Company or its subsidiaries.

Baricitinib (LY3009104) Atopic Dermatitis

Study I4V-MC-JAIY is a Phase 3, multicenter, randomized, double-blind, placebo-controlled, parallel-group, outpatient study evaluating the efficacy and safety of baricitinib 4 mg once daily (QD) plus topical corticosteroids (TCS) and 2 mg QD plus TCS, as compared to placebo plus TCS in adult patients with moderate to severe atopic dermatitis.

Eli Lilly and Company
Indianapolis, Indiana USA 46285
Protocol I4V-MC-JAIY
Phase 3
Statistical Analysis Plan Version 1 electronically signed and approved by Lilly:
01 April 01, 2019
Statistical Analysis Plan Version 2 electronically signed and approved by Lilly on date provided below.

Approval Date: 12-Aug-2019 GMT

2. Table of Contents

LY3009104

Table of Contents

Table   **Page**

Table of Contents

3.   Revision History

Statistical Analysis Plan (SAP) Version 1 is based on Protocol I4V-MC-JAIY(a). Statistical Analysis Plan Version 2 was approved prior to unblinding. A summary of changes between Version 1 and Version 2 are as follows:

Section	Summary of Changes
4.3. Exploratory objectives	• Updated the time frame for the objective of frequency of patient-reported "no itch" and "no pain" by starting from Week 0. • Added new exploratory analyses on Hospital Anxiety Depression (HADS), Dermatology Life Quality Index (DLQI), Atopic Dermatitis Sleep Scale (ADSS), and Patient-Oriented Eczema Measure (POEM). • Added new exploratory analyses on early responders.
6.11.1. Background TCS	• Updated the weight of dispensed topical corticosteroids (TCS) tube to align with JAHL SAP.
6.2.2. Definition on Baseline and Postbaseline Measures	• Updated the derivations of itch weekly scores.
6.2.3. Analysis Methods	• Updated the structures of covariance matrix in mixed model repeated measures (MMRM) model.
6.6. Multiple Comparisons	• Updated the graphical testing procedure.
6.16 Subgroup Analysis	• Added subgroup analyses for East Asia vs. other.
Table 6.6	• Updated the endpoints on the number days of itch-free and pain-free days.
Table 6.7	• Updated analyses on the number of itch-free and pain-free days. • Change the type of analysis of covariance (ANCOVA) (modified last observation carried forward [mLOCF]) analyses on PROMIS from "exploratory analyses" to "sensitivity analyses."

4. Study Objectives

4.1. Primary Objective

The primary objective of this study is to test the hypothesis that baricitinib 4 mg once daily (QD) plus topical corticosteroids (TCS) or baricitinib 2 mg QD plus TCS is superior to placebo plus TCS in the treatment of patients with moderate to severe atopic dermatitis (AD), as assessed by the proportion of patients achieving the validated Investigator's Global Assessment for AD (vIGA-AD, referred to throughout the SAP as IGA) of 0 or 1 with a ≥2-point improvement at Week 16.

In particular, the associated estimand for this objective is to measure the effect of therapy with baricitinib as assessed by the proportion of patients with a response of IGA 0 or 1 at Week 16 assuming treatment response disappears after patients are rescued or discontinue from the study or treatment. See Sections 6.4.1 and 6.12.1 on how this estimand handles outcomes after the occurrence of any intercurrent event through nonresponder imputation (NRI).

4.2. Secondary Objectives

4.2.1. Key Secondary Objectives

These are prespecified objectives that will be adjusted for multiplicity.

Objectives	Endpoints
To compare the efficacy of baricitinib 2 mg QD + TCS or baricitinib 4 mg QD + TCS to placebo + TCS in AD during the 16-week double-blind placebo-controlled treatment period as measured by improvement in signs and symptoms of AD.	• Proportion of patients achieving EASI75 at 16 weeks • Proportion of patients achieving EASI90 at 16 weeks • Percent change from baseline in EASI score at 16 weeks • Proportion of patients achieving SCORAD75 at 16 weeks
To compare the efficacy of baricitinib 2 mg QD + TCS or baricitinib 4 mg QD + TCS to placebo + TCS in AD during the 16-week double-blind placebo-controlled treatment period as assessed by patient-reported outcome measures.	• Proportion of patients achieving a 4-point improvement from baseline in Itch NRS at 2 days, 1 week, 2 weeks, 4 weeks, and 16 weeks • Mean change from baseline in the score of Item 2 of the ADSS at 1 week and 16 weeks • Mean change from baseline in Skin Pain NRS at 16 weeks

4.2.2. Other Secondary Objectives

These are prespecified objectives that will not be adjusted for multiplicity.

Objectives	Endpoints
To test the hypothesis that baricitinib 2 mg QD + TCS or baricitinib 4 mg QD + TCS is superior to placebo + TCS in the treatment of patients with moderate to severe AD.	• Proportion of patients achieving IGA of 0 or 1 with a ≥2-point improvement at Week 4
To compare the efficacy of baricitinib 2 mg QD + TCS or baricitinib 4 mg QD + TCS to placebo + TCS in AD during the 16-week double-blind placebo-controlled period as measured by signs and symptoms of AD.	• Proportion of patients achieving EASI50 at 16 weeks • Proportion of patients achieving IGA of 0 at 16 weeks • Mean change from baseline in SCORAD at 16 weeks • Proportion of patients achieving SCORAD90 at 16 weeks • Mean change from baseline in BSA affected at 16 weeks • Proportion of patients developing skin infections requiring antibiotic treatment by Week 16 • Mean gram quantity of background TCS used over 16 weeks (tube weights)
To compare the efficacy of baricitinib 2 mg QD + TCS or baricitinib 4 mg QD + TCS to placebo + TCS in AD during the 16-week, double-blind, placebo-controlled treatment period as assessed by patient-reported outcome/QoL measures.	• Percent change from baseline in Itch NRS at 2 days, 1 week, 4 weeks, and 16 weeks • Mean change from baseline in Itch NRS at 2 days, 1 week, 4 weeks and 16 weeks • Proportion of patients achieving a 4-point improvement from baseline in Skin Pain NRS at 16 weeks • Mean change from baseline in the total score of the POEM at 16 weeks • Mean change in the PGI-S-AD scores at 16 weeks • Mean change from baseline in HADS at 16 weeks • Mean change in the DLQI scores at 16 weeks • Mean change in the WPAI scores at 16 weeks • Mean change in the EQ-5D-5L scores at 16 weeks • Mean number of days without use of background TCS over 16 weeks

4.3. Exploratory Objectives

The exploratory objectives of this study are the following:

Objectives/Endpoints
• Frequency of patient-reported "no itch" (Itch NRS score = 0) days from daily diaries from Week 0 to Week 16
• Frequency of patient-reported "no pain" (Skin Pain NRS score = 0) days from daily diaries from Week 0 to Week 16
• Mean change from baseline in PIQ – Itch Interference score
• Mean change from baseline in PIQ – Activity and Clothing score
• Mean change from baseline in PIQ – Mood and Sleep score
• Mean change from baseline in PIQ – Scratching Behavior score
• Mean change from baseline in PROMIS – Sleep-Related Impairment score
• Mean change from baseline in Neuro-QoL – Cognitive Function score
• Patient Benefit Index score at 16 weeks – global score plus the following subscales:
o Reducing social impairments
o Reducing psychological impairments
o Reducing impairments due to therapy
o Reducing physical impairments
o Having confidence in healing
• Proportion of patients achieving PBI global score ≥1 at 16 weeks
• Mean change from baseline in the score of Item 1 of the ADSS at 1 week and 16 weeks
• Proportion of patients achieving a ≥1-point improvement in the score of Item 1 of the ADSS for those with baseline Item 1 score ≥1
• Proportion of patients achieving a ≥1-point improvement in the score of Item 2 of the ADSS for those with baseline Item 2 score ≥ 1
• Proportion of patients achieving a ≥2-point improvement in the score of Item 2 of the ADSS for those with baseline Item 2 score ≥2
• Mean change from baseline in the score of Item 3 of the ADSS at 1 week and 16 weeks
• Proportion of patients achieving a ≥1-point improvement in the score of Item 3 of the ADSS for those with baseline Item 3 score ≥1
• To evaluate changes from baseline in immunoglobulin E levels during the study
• To evaluate changes from baseline in eosinophil levels during the study
• To assess time to 4-point Itch NRS improvement during the first 14 days after initiation of treatment
• To assess time to 4-point improvement in Skin Pain during the first 14 days after the initiation of treatment
• Proportion of patients achieving a ≥4-point improvement in DLQI total score for those with baseline DLQI total score ≥4
• Proportion of patients achieving DLQI total score 0 or 1
• Proportion of patients achieving a ≥ 4-point improvement in POEM total score for those with baseline total score ≥4
• Proportion of patients achieving HADS Anxiety Score <8 for those with baseline HADS Anxiety Score ≥8
• Proportion of patients achieving HADS Depression Score <8 for those with baseline HADS Depression Score ≥8
• Proportion of patients achieving improvement with HADS Anxiety Score or HADS Depression Score <8 for those with baseline HADS Anxiety Score ≥8 or HADS Depression Score ≥8
• Mean change from baseline in HADS subscale scores

- Proportion of patients achieving ≥4-point improvement in Itch NRS or IGA≤1 at Week 16 for those with ≥3-point improvement in Itch NRS or IGA≤2 at Week 4
- Proportion of patients achieving ≥ 4-point improvement in Itch NRS or IGA≤1 at any time between Week 8 and Week 16 for those with ≥3-point improvement in Itch NRS or IGA≤2 at Week 4
- Proportion of patients achieving [≥4-point improvement in Itch NRS and IGA≤2] or [IGA≤1] at Week 16 for those with ≥3-point improvement in Itch NRS or IGA≤2 at Week 4
- Proportion of patients achieving [≥4-point improvement in Itch NRS and IGA≤2] or [IGA≤1] at any time between Week 8 and Week 16 for those with ≥3-point improvement in Itch NRS or IGA≤2 at Week 4

5. Study Design

5.1. Summary of Study Design

Study I4V-MC-JAIY (JAIY) is a Phase 3, multicenter, randomized, double-blind, placebo-controlled, parallel-group, outpatient study evaluating the efficacy and safety of baricitinib 2 mg QD and 4 mg QD, in combination with TCS, as compared to placebo in combination with TCS, in adult patients with moderate to severe AD. The study is divided into 3 periods, a 5-week Screening period, a 16-week Double-Blinded Treatment period, and a 4-week Post-Treatment Follow-Up period. For those patients who complete the 16-week treatment period, there is an option to participate in the long-term extension study I4V-MC-JAHN (JAHN).

Approximately 300 patients ≥18 years of age who have responded inadequately to topical therapy will be randomized at a 1:1:1 ratio to receive placebo QD, baricitinib 2 mg QD, or baricitinib 4 mg QD in combination with TCS (100 patients in each treatment group). Patients will be stratified at randomization according to disease severity (Investigator's Global Assessment [IGA] 3 vs. 4) and geographic region.

Study JAIY will consist of 3 periods:

- Period 1: Screening period is between 8 and 35 days prior to Week 0 (Visit 2)
- Period 2: Double-Blind, Placebo-Controlled Treatment period from Week 0 (Visit 2) through Week 16 (Visit 8)
- Period 3: Post-Treatment Follow-Up period from last treatment visit at Week 16 (Visit 8) or Early Termination Visit (ETV) to approximately 28 days after the last dose of investigational product

Figure JAIY.5.1 illustrates the study design. The blinding procedure is described in the Protocol.

Figure JAIY.5.1.   Illustration of study design for Clinical Protocol I4V-MC-JAIY.

Abbreviations: AD = atopic dermatitis; eGFR = estimated glomerular filtration rate; PPD = purified protein derivative; QD = once daily; TCS = topical corticosteroids; V = visit; W = week.

a Applicable to patients taking topical treatments (excluding emollients) or systemic treatments for AD at the time of screening.

b For patients randomized to the 4-mg once daily dose who have renal impairment (defined as eGFR <60 mL/min/1.73 m^2), the baricitinib dose will be 2 mg once daily.

c Patients for whom PPD skin test for the evaluation of tuberculosis infection was performed at V1 must return and PPD test must be read 48 to 72 hours after Visit 1 (post-PPD).

d Occurs approximately 28 days after the last dose of investigational product. Not required for those patients entering the long-term extension study JAHN.

5.2.   Method of Assignment to Treatment

Patients who meet all criteria for enrollment will be randomized in a 1:1:1 ratio (placebo; baricitinib 2 mg; baricitinib 4 mg) to double-blind treatment at Visit 2 (Week 0). Randomization will be stratified by geographic region (Europe [EU], Japan [JPN], rest-of-world [ROW]) and disease severity at baseline (IGA 3 vs. 4). Assignment to treatment groups will be determined by a computer-generated random sequence using an IWRS. The IWRS will be used to assign packages containing double-blind investigational product tablets to each patient according to the study schedule of activities. Site personnel will confirm that they have located the correct packages by entering a confirmation number found on the packages into the IWRS.

This study will be conducted internationally in multiple sites. Table JAIY.5.1 describes how regions will be defined for stratification. Regions may be combined for statistical analyses in the

case when one of the region strata fails to meet the required minimum number of 30 patients. The 2 region strata with the least number of patients will then be pooled.

Table JAIY.5.1. **Geographic Regions for Stratification**

Region	Countries
Europe	Germany, Italy, Poland, Spain, Austria
Japan	Japan
Rest of World	Taiwan, Australia, Korea, Argentina

6. A Priori Statistical Methods

6.1. Determination of Sample Size

Study JAIY will aim to enroll approximately 300 patients ≥18 years of age. The proposed sample size will ensure a 89% power to detect an absolute difference of 20% between the baricitinib 4 mg and placebo treatment groups and the baricitinib 2 mg and placebo treatment groups, each using a 2-sided alpha of 0.025 and a Fisher′s exact test, assuming a 10% placebo response rate for the primary endpoint. The assumptions are based on what was observed in the Phase 2 study (JAHG). The proposed end point of IGA 0 or 1 represents patients whose AD is clear or almost clear from a baseline of moderate or severe disease. The anticipated effect size represents 2 times more patients achieving this benefit compared to placebo, which, in discussion with therapeutic experts, is of a magnitude that is considered clinically relevant.

Sample size and power estimates were obtained from nQuery® Advisor 7.0.

6.2. General Considerations

This plan describes *a priori* statistical analyses for efficacy, health outcomes, and safety that will be performed.

Statistical analysis of this study will be the responsibility of Eli Lilly and Company (Lilly). The statistical analyses will be performed using SAS® Version 9.4 or higher.

Not all displays described in this SAP will necessarily be included in the clinical study report (CSR). Not all displays will necessarily be created as a "static" display. Some may be incorporated into interactive display tools instead of or in addition to a static display. Any display described in this SAP and not included in the CSR will be available upon request.

Statistical tests of treatment effects and confidence intervals (CIs) will be performed at a 2-sided significance level of 0.05, unless otherwise stated (e.g., graphical multiple testing strategy in Section 6.6).

Data collected at ETVs will be mapped to the closest scheduled visit number for that patient if it falls within the visit window as discussed in Section 6.2.2. For by-visit summaries, only visits in which a measure was scheduled to be collected will be summarized. Any unscheduled visit data will be included at the patient-level listings. However, the data will still be used in other analyses, including shift analyses for safety analytes, change from baseline using modified last observation carried forward (mLOCF) for efficacy analyses, and other categorical analyses including safety.

6.2.1. Analysis Populations

Intent-to-treat (ITT) population: The ITT population analysis set is defined as all randomized patients.

Per-protocol Set (PPS): PP subset of the ITT analysis set will include those patients who do not have any significant or important protocol violations. Qualifications for and identification of

significant or important protocol violations will be determined while the study remains blinded, prior to database lock.

Follow-up population: The follow-up population is defined as patients who entered the follow-up period.

Unless otherwise specified, the efficacy and health outcome analyses will be conducted on the ITT population (Gillings and Koch 1991), which seeks to preserve the benefits of randomization and avoid selection bias. Patients will be analyzed according to the treatment to which they were randomized. In addition, the analyses of primary and key secondary endpoints will be repeated using the PPS population.

Safety population: The safety population is defined as all randomized patients who receive at least 1 dose of investigational product and who did not discontinue from the study for the reason 'Lost to Follow-up' at the first postbaseline visit.

Safety analyses will be performed using the safety population. Patients will be analyzed according to the treatment regimen to which they were assigned. Analyses of the safety endpoints, many of which are incidence based, will include all patients in the safety population, unless specifically stated otherwise.

In the rare situation where a patient is Lost to Follow-up at the first postbaseline visit, but some safety data exists (e.g., unscheduled laboratory assessments) after first dose of study drug, a listing of the data or a patient profile may be provided, when requested.

6.2.2. *Definition of Baseline and Postbaseline Measures*

The baseline value for efficacy and health outcomes variables measured at scheduled visits is defined as the last non-missing measurement on or prior to the date of first study drug administration (expected at Week 0, Visit 2).

The baseline value for the daily diary assessments (Itch NRS, ADSS, Skin Pain NRS, PGI-S-AD) is the mean of the non-missing assessments in the 7 days prior to the date of first study drug administration (expected at Week 0, Visit 2).

If there are less than 4 non-missing assessments in the baseline diary window, the interval lower bound can be extended up to 7 additional days, one day at a time, to obtain the most recent 4 non-missing values. If there are not at least 4 non-missing assessments in the baseline period, the baseline mean is missing.

Baseline for the safety analyses is defined as the last non-missing scheduled (planned) measurement on or prior to the date of first study drug administration for continuous measures by-visit analyses and all non-missing measurements on or prior to the date of first study drug administration for all other analyses.

Postbaseline measurements are collected after study drug administration through Week 16 (Visit 8) or early discontinuation visit. Efficacy data collected at scheduled visits (e.g., eCOA, ClinRO) will be used in all analyses unless it is missing. If an assessment is missing at a scheduled visit, an unscheduled post-baseline assessment can be used provided it falls within the

window interval as follows: a ±2 day window is used for Visit 3 (Week 1), Visit 4 (Week 2), and Visit 5 (Week 4); and a ±4 day window is used to Visit 6 (Week 8), Visit 7 (Week 12), and Visit 8 (Week 16). If there is more than 1 unscheduled visit within the defined visit window and no scheduled visit assessment is available, the unscheduled visit closest to the scheduled visit date will be used. If two unscheduled visits of equal distance are available, then the latter of the two will be used.

Postbaseline daily diary endpoints will be the mean of weekly visit windows (diary windows) anchored on day of first dose (Day 1) for Weeks 1 through 14 as follows: Week 1 (Days 1 through7), Week 2 (Days 8 through 14), Week 3 (Days 15 through 21), …, Week 14 (Days 92 through 98).

Week 16 Daily Diary Window Construction

The following sequential steps will be used to determine the Week 16 diary window. The general goal is to anchor on the scheduled Week 16 visit (or a proximal unscheduled visit) if such a visit exists or to use an interval based on days in study for cases where a scheduled Week 16 or a proximal surrogate does not exist.

Step 1: If the Week 16 scheduled visit exists, the Week 16 diary interval is the 7 days prior to the Week 16 date provided that window has at least 4 non-missing observations. If there are less than 4 non-missing observations, the diary window's lower bound will be extended 1 day at a time (up to day 99) to a maximum of 14 days prior to the Week 16 date until 4 non-missing observations are obtained. If, after extending this diary window's lower bound to 14 days, there are less than 4 non-missing observations then go to Step 2.

Step 2: If the Week 16 scheduled visit does not exist, the 7 days prior to the last visit (scheduled or unscheduled) occurring after Day 105, will constitute the Week 16 diary window provided that window contains at least 4 non-missing observations. If there are less than 4 non-missing observations, the diary window's lower bound will be extended 1 day at a time (up to Day 99) to a maximum of 14 days prior to the unscheduled visit date until 4 non-missing observations are obtained. If, after extending this diary window's lower bound to 14 days, there are less than 4 non-missing observations then go to Step 3.

Step 3: If neither a Week 16 scheduled visit is available nor an unscheduled visit to act as a surrogate for the Week 16 diary window, then the Week 16 window will be Day 106 to Day 112. If there are less than 4 non-missing observations, the dairy window's lower bound will be extended 1 day at a time to Day 99 until 4 non-missing observations are obtained.

If the steps above do not detect a window with at least 4 non-missing observations, then the Week 16 window is 7 days from either the Week 16 visit, the surrogate visit or Days 106 through 112 and the mean is missing and subject to imputation rules.

Week 15 Daily Diary Window Construction

The lower boundary of the Week 15 diary window is defined as Day 99. The upper bound of the Week 15 diary window is the minimum of either Day 105 or the lower bound of the Week 16

diary window -1. Consequently, Week 15 may be less than 4 days if the Week 16 scheduled visit is before Day 112. Moreover, as Week 15 diary window cannot exceed 7 days, there could be daily assessments between Weeks 15 and 16 diary windows that do not fall into a diary window. If after constructing the diary windows, there are fewer than 4 non-missing values the mean for Week 15 is missing and subject to imputation rules.

Handling of Duplicate Diary Records

If there is more than one diary record on a particular date, the first record on that particular date will be used in the analysis.

Note, as some analyses require use of the primary censoring rule, assessments collected on the day of rescue or afterwards will be excluded from the weekly visit interval calculation when implementing the rule for daily diary. If, after exclusion of these records, there are less than 4 non-missing assessments, the weekly interval which implements the primary censoring rule will be missing. The post-study follow-up weekly score for daily diaries will be calculated as the mean of the 7 days prior to the follow-up visit which occur after last dose of study treatment.

Postbaseline measures for the safety analyses are defined as the non-missing scheduled (planned) measurements after the date of first study drug administration for continuous measures by-visit analyses and all non-missing measurements after the date of first study drug administration for all other analyses.

6.2.3. Analysis Methods

The main analysis method of categorical efficacy variables and health outcomes variables will use a logistic regression analysis with region, baseline disease severity (IGA), baseline value and treatment group in the model, except for the analysis on IGA. For IGA, the logistic regression model will include region, baseline disease severity (IGA), and treatment group. Firth's correction will be used in order to accommodate (potential) sparse response rates. The p-value for the odds ratio from the logistic regression model will be used for statistical inference, unless Firth's correction still results in quasi-separation. In that case, Fisher's exact test will be used for statistical inference. The difference in percentages and 100(1-alpha)% CI of the difference in percentages using the Newcombe-Wilson method without continuity correction will be reported. The p-value from the Fisher's exact test will also be produced as a secondary analysis.

The main analysis method for all continuous efficacy and health outcomes variables will use mixed model repeated measures (MMRM) analysis. The MMRM model will use a restricted maximum likelihood (REML) estimation. The model will include treatment, region, baseline disease severity (IGA), visit, and treatment-by-visit-interaction as fixed categorical effects and baseline and baseline-by-visit-interaction as fixed continuous effects. For daily diary assessments, the model for analyses up to Week 16 will include all weekly assessments. An unstructured (co)variance structure will be used to model the between- and within-patient errors. If this analysis fails to converge, the heterogeneous autoregressive [ARH(1)], followed by the heterogeneous compound symmetry (CSH), followed by the heterogeneous Toeplitz (TOEPH), followed by autoregressive [AR(1)], followed by compound symmetry (CS) will be used. The

Kenward-Roger method will be used to estimate the degrees of freedom. Treatment least squares means (LSM) will be estimated within the framework of the MMRM using type 3 sums of squares. Differences in LSM between each dose of baricitinib and placebo (and associated p-values, standard errors and 95% CI) will be used for statistical inference. The LSM difference, standard error, p-value and 95% CI will be reported.

Treatment comparisons for continuous efficacy and health outcomes variables may also be made using analysis of covariance (ANCOVA) for primary and key secondary objectives. When an ANCOVA model is used, the model includes region, baseline disease severity, treatment group, and baseline value. Treatment LSM will be estimated within the framework of the ANCOVA using type 3 sums of squares. Reported differences in LSM and associated p-values, standard errors and 95% CI will be used for statistical inference. Treatment-by-region interaction will also be added to the model for sensitivity purposes and is discussed in Section 6.5.

Beginning on Day 14 a Cox proportional hazard (CPH) model of time (in days) to first observance of a 4-point itch reduction with effects for treatment, region, baseline mean itch, and disease severity will be used to test for treatment differences from placebo. For this analysis, daily itch scores will be compared to the baseline to determine if a 4-point itch reduction has been achieved in patients with a baseline itch of at least 4. The baseline for Itch NRS is defined in Section 6.2.2. Beginning on Day 14, the day on which the first time itch NRS is reduced by at least 4 will be modeled. If any significant difference between any baricitinib dose and placebo is observed then the same analyses will be run on Day 13. This process of evaluating at the next lowest day will proceed until no significant differences are observed. No adjustments for multiple tests and multiple comparisons will be used. This analysis uses the Primary Censoring Rule for patients who are rescued or permanently discontinue study drug (see Section 6.4). Missing daily itch data will be replaced using NRI rule which means missing data is replaced with a non-response which would entail replacing missing values with a time to event of >14 days, censored; >13 days, censored, etc., depending on the window being used. If the model assumptions for the CPH model do not hold, a log-rank test will be used.

Restricted mean survival time will be evaluated at Days 2, 3, 4 and 5 as an exploratory analysis. The model will include terms for baseline disease severity (IGA), baseline itch, region, and treatment group.

Fisher's exact test will be used to test for differences between each baricitinib dose and placebo in proportions of patients experiencing adverse events (AEs), discontinuation from study drug, and for other categorical safety data. Continuous vital signs, body weight, and other continuous safety variables, including laboratory variables will be analyzed by an ANCOVA with treatment group and baseline value in the model. The significance of within-treatment group changes from baseline will be evaluated by testing whether or not the treatment group LSM changes from baseline are different from zero; the standard error for the LSM change will also be displayed. Differences in LSM will be displayed, with the p-value associated with the LSM comparison to placebo and a 95% CI on the LSM difference also provided. In addition to the LSMs for each group, the within-group p-value for the change from baseline will be displayed.

Confusion table will be used to calculate the proportions of responder patients defined at week 16, and at any time between Weeks 8 and 16, for those with improvement on itch NRS or IGA at early week. The confusion table will generate the Positive Prediction Value (PPV) and Negative Prediction Value (NPV), which can access the performance of predicting late week responders using early week data. All missing values will be treated by NRI method.

6.2.4. Derived Data

- Age (year), derived using first dose date as the reference start date and July 1st of birth year and truncated to a whole-year (integer) age. Patients whose derived age is less than 18 will have the required minimum age of 18 at informed confirmed; reporting for age, age groups, and lab ranges, however will be based on their derived age.
- Age group (<65, ≥65 years old)
- Age group (<65, ≥65 to <75, ≥75 to <85, ≥85 years old)
- Body Mass Index (BMI) (kg/m^2) = Weight (kg)/((Height (cm)/100)^2)
- BMI category (<25 kg/m^2, ≥25 to <30 kg/m^2, ≥30 kg/m^2)
- The duration of AD from diagnosis (years) = [(Date of informed consent – Date of AD diagnosis)+1]/ 365.25.
    - If year of onset is missing, duration of AD will be set as missing. Otherwise, unknown month will be taken as January, and unknown day will be taken as 01. The duration of AD will be rounded to 1 decimal place.
- Duration of AD (years) category (0 to <2 years, 2 to < 5 years, 5 to <10 years, 10 to <20 years, ≥20 years)
- Diagnosis age (years), derived using diagnosis date as the reference start date and July 1 of birth year and truncated to a whole-integer age.
- Diagnosis age group (<18, ≥18 to <50, ≥50 years old)
- Change from baseline = postbaseline measurement at Visit x – baseline measurement.
    - If a baseline value is missing, it will not be imputed and the change from baseline will not be calculated.
- Percent change from baseline at Visit x:
  ((Post-baseline measurement at Visit x - Baseline measurement)/Baseline measurement)*100.
    - If a baseline value is missing, it will not be imputed and percent change from baseline will not be calculated.
- Weight (kg) = weight (lbs) * 0.454.
- Weight category (<60 kg, ≥60 to <100 kg, ≥100 kg)
- Height (cm) = height (in) * 2.54.
- Cyclosporine inadequate efficacy (yes, no)
    - Set **yes** if the reason for discontinuation is inadequate response.
- Cyclosporine intolerance (yes, no)
    - Set **yes** if the reasons for discontinuation are: intolerance to medication or contraindication (Physician indicated cyclosporine was used and a contraindication was noted).
- Cyclosporine contraindication [ineligible] (yes, no)

- o Set to **yes** if cyclosporine never used because of a contraindication
- Cyclosporine inadvisable (yes, no)
  - Set to **yes** if the following reasons were selected for either not using the medication or discontinuing the medication:
    - o Reason for not using medication: Physician decision, concern about side effects, unfavorable benefit risk, contraindication.
    - o Reasons for discontinuation: inadequate response, intolerance to medication, or contraindication.
- TCNI inadequate efficacy (yes, no)
  - o Set **yes** if the reason for discontinuation is inadequate response.
- TCNI intolerance (yes, no)
  - o Set **yes** if the reasons for discontinuation are: intolerance to medication or contraindication (Physician indicated TCNI was used and a contraindication was noted).
- TCNI contraindication / [ineligible](yes, no)
  - o Set to **yes** if TCNI never used because of a contraindication
- TCNI inadvisable (yes, no)
  - Set to **yes** if the following reasons were selected for either not using the medication or discontinuing the medication:
    - o Reason for not using medication: Physician decision, concern about side effects, unfavorable benefit risk, contraindication.
    - o Reasons for discontinuation: inadequate response, intolerance to medication, or contraindication.

6.3. Covariate Adjustment

The randomization to treatment groups at Week 0 (Visit 2) is stratified by disease severity (IGA) and geographic region as described in Section 5.1. Unless otherwise specified, the statistical analysis models will adjust for these stratification variables. The covariates used in the logistic model for categorical data will include the parameter value at baseline. The covariates used in the ANCOVA model for continuous data will include the parameter value at baseline. Inclusion of baseline in the model ensures treatment LSM are estimated at the same baseline value. When an MMRM analysis is performed, baseline value and baseline-by-visit interactions will be included as covariates.

6.4. Handling of Dropouts or Missing Data

Intercurrent events (ICH E9 R1) are events which occur after the treatment initiation and make it impossible to measure a variable or influence how it would be interpreted.

Depending on the estimand being addressed, different methods will be used to handle missing data as a result of intercurrent events. Intercurrent events can occur through the following:

- application of one of the censoring rules (including after permanent study drug discontinuation or after rescue therapy)
- discontinuation

- missing an intermediate visit prior to discontinuation or rescue
- lost to follow-up.

Non-censor intercurrent events are events that are not due to the application of any censoring rule, i.e., the last 3 items in the list above.

Note that as efficacy and health outcome data can accrue after a patient permanently discontinues study drug or begins rescue therapy, specific general censoring rules to the data will be applied to all efficacy and health outcome observations subsequent to these events depending on the estimand being addressed. These specific censoring rules are described below.

The *primary censoring rule* will censor efficacy and health outcome data after permanent study drug discontinuation or after rescue therapy. This censoring rule will be applied to all continuous and categorical efficacy and health outcome endpoints. This censoring rule is equivalent to using all the data up to rescue.

A *secondary censoring rule* will only censor efficacy and health outcome data after permanent study drug discontinuation. This sensitivity analysis will include all observed values up to study drug discontinuation. The secondary censoring rule will be applied to primary and key secondary efficacy and health outcome endpoints as sensitivity analyses.

Table JAIY.6.1 describes the planned imputation methods for efficacy and health outcome endpoints with associated censoring rules. Sections 6.4.1 through 6.4.5 summarize the methodology of each imputation rule.

Table JAIY.6.1.   Imputation Techniques for Various Variables

Efficacy and Health Outcome Endpoints	Imputation Method
IGA(0,1), EASI75, 4-point Itch NRS improvement, EASI90, SCORAD75	NRI[a][b], pMI[a], Tipping point[a]
EASI percent change, ADSS Item 2 change, Skin Pain NRS change	MMRM[a][b], mLOCF[a], pMI[a]
All remaining categorical measures	NRI[a]
All remaining continuous efficacy and health outcome measures	MMRM[a], mLOCF[a]

Abbreviations: ADSS = Atopic Dermatitis Sleep Scale; EASI = Eczema Area and Severity Index score; IGA = Investigator's Global Assessment for AD; mLOCF = modified last observation carried forward; MMRM = mixed model repeated measures; NRI = nonresponder imputation, NRS = Numeric Rating Scale; pMI = placebo multiple imputation; SCORAD = SCORing Atopic Dermatitis.
[a] Analyses utilizing the primary censoring rule.
[b] Analyses utilizing the secondary censoring rule.

6.4.1. *Nonresponder Imputation*

A nonresponder imputation (NRI) method imputes missing values as non-responses and can be justified based on the composite strategy for handling intercurrent events (ICH E9 R1). This imputation procedure assumes the effects of treatments disappear after the occurrence of an intercurrent event defined by the associated censoring rule.

All categorical endpoints will utilize the NRI method after applying the primary censoring rule to patients who permanently discontinued study drug or were rescued (described in Section 6.4). Additionally, all primary and key secondary categorical endpoints will utilize NRI after applying the secondary censoring rule as sensitivity analyses. For analyses which utilize either of the censoring methods, randomized patients without at least 1 post-baseline observation will be defined as nonresponders for all visits.

6.4.2. *Mixed Model for Repeated Measures*

Mixed Model for Repeated Measures analyses will be performed on continuous endpoints to mitigate the impact of missing data. This approach assumes missing observations are missing-at-random (missingness is related to observed data) and borrows information from patients in the same treatment arm taking into account both the missingness of data through the correlation of the repeated measurements.

Essentially MMRM estimates the treatment effects had all patients remained on their initial treatment throughout the study. For this reason, the MMRM implies a different estimand (hypothetical strategy [ICH E9 R1]) than the one used for NRI on categorical outcomes.

All continuous endpoints will utilize MMRM after applying the primary censoring rule. As sensitivity analyses, all secondary continuous endpoints will also utilize MMRM after applying the secondary censoring rule (Table JAIY.6.1).

6.4.3. *Modified Last Observation Carried Forward*

For continuous measure, a modified last observation carried forward (mLOCF) imputation technique replaces missing data with the most recent non-missing post-baseline assessment. The specific modification to the LOCF is data after an intercurrent event will not be carried forward thus the mLOCF is applied after the specified censoring rule is implemented. The mLOCF assumes the effect of treatment remain the same after the event that caused missing data as it was just prior to the missing data event. Analyses using mLOCF require a nonmissing baseline and at least 1 postbaseline measure otherwise the data is missing for analyses purposes. Analyses using mLOCF help ensure the number of randomized patients who were assessed post-baseline is maximized and is reasonable for this data as data directly prior to an intercurrent event (such as initiation of rescue therapy or drop out) is likely a non-efficacious response.

All continuous efficacy and health outcomes endpoints will use with mLOCF imputation methodology with an ANCOVA as sensitivity analyses to the MMRM analyses.

6.4.4. *Placebo Multiple Imputation*

The Placebo Multiple Imputation (pMI) methodology will be used as a sensitivity analysis for the analysis of the primary efficacy endpoint (IGA 0 or 1 at Week 16) as well as the key secondary endpoints at Week 16. In these sensitivity analyses the primary censoring rule will be applied.

The pMI assumes that the statistical behavior of drug- and placebo- treated patients after the occurrence of intercurrent events will be the same as if patients were treated with placebo. Thus,

in the effectiveness context, pMI assumes no pharmacological benefit of the drug after the occurrence of intercurrent events but is a more conservative approach than mLOCF because it accounts for uncertainty of imputation, and therefore does not underestimate standard errors, and it limits bias. In the efficacy context pMI is a specific form of a missing not at random analysis and expected to yield a conservative estimate of efficacy.

In the pMI analysis, multiple imputations are used to replace missing outcomes for drug- and placebo-treated patients who have an intercurrent event using multiple draws from the posterior predictive distribution estimated from the placebo arm. The binary outcomes will then be derived from the imputed data.

Data are processed sequentially by repeatedly calling SAS® PROC MI to impute missing outcomes at visits $t=1,..,$ T.

1. *Initialization:* Set $t=0$ (baseline visit)

2. *Iteration:* Set $t=t+1$. Create a data set combining records from drug- and placebo-treated patients with columns for covariates **X** and outcomes at visits $1,..,t$ with outcomes for all drug-treated patients set to missing at visit t and set to observed or imputed values at visits $1,..,t-1$.

3. *Imputation:* Run Bayesian regression in SAS® PROC MI on this data to impute missing values for visit t using previous outcomes for visits 1 to $t-1$ and baseline covariates. Note that only placebo data will be used to estimate the imputation model since no outcome is available for drug-treated patients at visit t.

4. Replace imputed data for all drug-treated patients at visit t with their observed values, whenever available up to permanent study drug discontinuation and/or rescue (if censoring on rescue). If $t < $ T then go to Step 2, otherwise proceed to Step 5.

5. Repeat steps 1-4, m times with different seed values to create m imputed complete data sets.

Analysis: For continuous endpoints, fit its treatment response model (MMRM) for each completed data set. For the primary and secondary key efficacy endpoints [IGA (0,1), EASI75, EASI90, SCORAD75, and 4-point improvement from baseline in Itch NRS], the binary outcomes will be derived from the imputed data for each patient before fitting the logistic regression model.

The number of imputed data sets will be $m=100$ and a 6-digit seed value will be pre-specified for each analysis. Within the program, the seed will be used to generate the m seeds needed for imputation. The initial seed values are given in Table JAIY.6.2.

Table JAIY.6.2.   **Seed Values for Multiple Imputation**

Analysis	Seed value
Proportion of patients achieving IGA of 0 or 1 with a ≥2-point improvement from baseline at Week 16 using the primary censoring rule	123450
Percent change from baseline in EASI score at 16 weeks using the primary censoring rule. EASI75 and EASI90 will leverage imputation from EASI and therefore do not need a new seed number.	123451
Proportion of patients achieving SCORAD75 at 16 week using the primary censoring rule, with data up to rescue	123452
Proportions of patients achieving a 4-point improvement from baseline in Itch NRS at Week 16 using the primary censoring rule	123453
Mean change from baseline in Skin Pain NRS at Week 16 using the primary censoring rule	123454
Mean change from baseline in the score of Item 2 of the ADSS at Week 16 using the primary censoring rule	123455

Abbreviations: ADSS = Atopic Dermatitis Sleep Scale; EASI = Eczema Area and Severity Index score;
   IGA = Investigator's Global Assessment for AD; NRS = Numeric Rating Scale; SCORAD = SCORing Atopic
   Dermatitis.

The final inference on treatment difference is conducted from the multiple datasets using Rubin's combining rules, as implemented in SAS® PROC MIANALYZE.

6.4.5. Tipping Point Analyses

To investigate the missing data mechanism, sensitivity analyses using multiple imputation (MI) under the missing not at random assumption will be provided for the following primary and key secondary objectives as given in Table JAIY.6.3.

All patients in the ITT population will be included. Data after the occurrence of intercurrent events (after application of the primary censoring rule) will be set to missing.

Within each analysis, a most extreme case will be considered, in which all missing data for patients randomized to baricitinib 2 mg or 4 mg will be imputed using the worst possible result and all missing data for patients randomized to placebo will be imputed with the best possible result. Treatment differences will be analyzed using logistic regression or ANCOVA (Section 6.1) as appropriate.

For continuous variables, the following process will be used to determine the tipping point:

1. To handle intermittent missing visit data, a Markov chain Monte Carlo method (SAS® Proc MI with MCMC option) will be used to create a monotone missing pattern.

2. A set of Bayesian regressions (using SAS® Proc MI with MONOTONE option) will be used for the imputation of monotone dropouts. Starting from the first visit with at least 1 missing value, the regression models will be fit sequentially with treatment as a fixed effect and values from the previous visits as covariates.

3. A delta score is added to all imputed scores at the primary time point for patients in the baricitinib treatment groups, thus worsening the imputed value. The delta score is capped for patients based on the range of the outcome measure being analyzed.

4. Treatment differences between baricitinib and placebo are analyzed for each imputed dataset using ANCOVA (Section 6.1). Results across the imputed datasets are aggregated using SAS® Proc MIANALYZE in order to compute a p-value for the treatment comparisons for the given delta value.

5. Steps 3 and 4 are repeated, and the delta value added to the imputed baricitinib scores is gradually increased. The tipping point is identified as the delta value which leads to a loss of statistical significance (aggregated p-value >0.05) when evaluating baricitinib relative to the placebo group.

As a reference, for each delta value used in Steps 3 through 5, a fixed selection of delta values (ranging from slightly negative to slightly positive) will be added to imputed values in the placebo group, and Step 4 will be performed for the combination. This will result in a 2-d table, with the columns representing the delta values added to the imputed placebo responses, and the rows representing the delta values added to the imputed baricitinib responses. Separate 2-d tables will compare each baricitinib dose group to placebo.

A similar process will be used for the categorical variables:

1. Missing responses in the baricitinib groups will be imputed with a range of low response probabilities, including probabilities of 0, 0.1, and 0.2.

2. For missing responses in the placebo group, a range of responses probabilities (for example, probability = 0, 0.2 … 1) will be used to impute the missing values. Multiple imputed datasets will be generated for each response probability.

3. Treatment differences between baricitinib and placebo are analyzed for each imputed dataset using logistic regression (Section 6.1). Results across the imputed datasets are aggregated using SAS® Proc MIANALYZE in order to compute a p-value for the treatment comparisons for the given response probability. If the probability values do not allow for any variation between the multiple imputed datasets (for example, all missing responses in the placebo and baricitinib groups are imputed as responders and nonresponders, respectively), then the p-value from the single imputed dataset will be used.

The tipping point is identified as the response probability value within the placebo group that leads to a loss of statistical significance when evaluating baricitinib relative to placebo.

For tipping point analyses the number of imputed data sets will be m=100 and the seed values to start the pseudorandom number generator of SAS Proc MI (same values for MCMC option and for MONOTONE option) are given in Table JAIY.6.3.

Table JAIY.6.3.   Seed Values for Imputation

Analysis	Seed value
Proportion of patients achieving IGA (0,1) with ≥2-point improvement at Week 16; primary censoring rule	123470
Proportion of patients achieving EASI75 at Week 16; primary censoring rule Proportion of patients achieving EASI90 at Week 16; primary censoring rule	123471
Proportions of patients achieving a 4-point improvement from baseline in Itch NRS at Week 16, primary censoring rule	123472
Proportion of patients achieving SCORAD75 at Week 16; primary censoring rule	123473

6.5. Multicenter Studies

This study will be conducted by multiple investigators at multiple sites internationally. The countries will be categorized into geographic regions, as described in Section 5.2.

For the analysis of the primary endpoint, treatment-by-region interaction will be added to the logistic regression model as a sensitivity analysis and results from this model will be compared to the primary model (without the interaction effect). If the treatment-by-region interaction is significant at a 2-sided α level of 0.1, the nature of this interaction will be inspected as to whether it is quantitative (i.e., the treatment effect is consistent in direction across all regions but not in size of treatment effect) or qualitative (the treatment is beneficial in some but not all regions). If the treatment-by-region interaction effect is found to be quantitative, results from the primary model will be presented. If the treatment-by-region interaction effect is found to be qualitative, further inspection will be used to identify in which regions baricitinib is found to be more beneficial.

6.6. Multiple Comparisons/Multiplicity

The primary and key secondary endpoints will be adjusted for multiplicity in order to control the overall family-wise Type I error rate at a 2-sided alpha level of 0.05.

The following is a list of primary and key secondary endpoints to be tested.

Primary Null Hypotheses:

- Null Hypotheses[IGA0-1]: Proportion of baricitinib 4-mg patients achieving IGA of 0 or 1 with a ≥2-point improvement from baseline at Week 16 is equal to the proportion of placebo patients achieving IGA of 0 or 1 with a ≥2-point improvement from baseline at Week 16
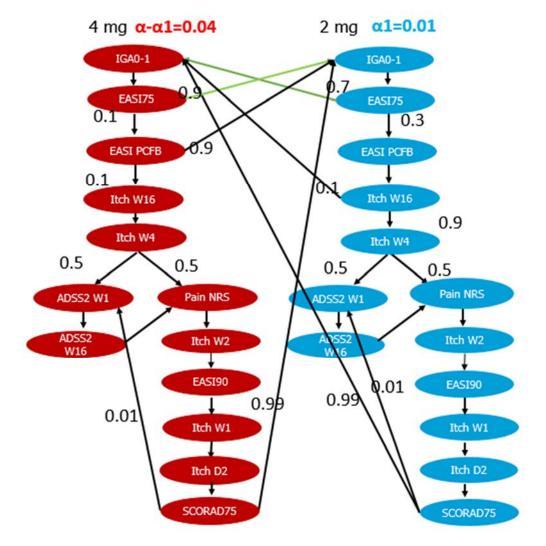- Null Hypotheses[IGA0-1]: Proportion of baricitinib 2-mg patients achieving IGA of 0 or 1 with a ≥2-point improvement from baseline at Week 16 is equal to the proportion of placebo patients achieving IGA of 0 or 1 with a ≥2-point improvement from baseline at Week 16

Key Secondary Null Hypotheses:
- Null Hypotheses[EASI75]: Proportion of baricitinib 4-mg patients achieving EASI75 is equal to the proportion of placebo patients achieving EASI75 at Week 16
- Null Hypotheses[ITCH W16]: Proportion of baricitinib 4-mg patients achieving a 4-point improvement in Itch NRS is equal to the proportion of placebo patients achieving a 4-point improvement in Itch NRS at Week 16 among patients with baseline Itch NRS score ≥4
- Null Hypotheses[EASI PCFB]: Percent change from baseline in EASI score for baricitinib 4-mg patients is equal to the percent change from baseline in EASI score for placebo patients at Week 16
- Null Hypotheses[ITCH W4]: Proportion of baricitinib 4-mg patients achieving a 4-point improvement in Itch NRS is equal to the proportion of placebo patients achieving a 4-point improvement in Itch NRS at Week 4 among patients with baseline Itch NRS score ≥4
- Null Hypotheses[SCORAD75]: Proportion of baricitinib 4-mg patients achieving SCORAD75 is equal to the proportion of placebo patients achieving SCORAD75 at Week 16
- Null Hypotheses[EASI 90]: Proportion of baricitinib 4-mg patients achieving EASI90 is equal to the proportion of placebo patients achieving EASI90 at Week 16
- Null Hypotheses[PAIN NRS]: Mean change from baseline in Skin Pain NRS for baricitinib 4-mg patients is equal to the mean change from baseline in Skin Pain NRS for placebo patients at Week 16
- Null Hypotheses[ADSS2 W16]: Mean change from baseline in the score of Item 2 of the ADSS for baricitinib 4-mg patients equal to the mean change from baseline in the score of Item 2 of the ADSS for placebo patients at Week 16
- Null Hypotheses[ITCH W2]: Proportion of baricitinib 4-mg patients achieving a 4-point improvement in Itch NRS is equal to the proportion of placebo patients achieving a 4-point improvement in Itch NRS at Week 2 among patients with baseline Itch NRS score ≥4
- Null Hypotheses[ADSS2 W1]: Mean change from baseline in the score of Item 2 of the ADSS for baricitinib 4-mg patients is equal to the mean change from baseline in the score of Item 2 of the ADSS for placebo patients at Week 1
- Null Hypotheses[ITCH W1]: Proportion of baricitinib 4-mg patients achieving a 4-point improvement in Itch NRS is equal to the proportion of placebo patients achieving a 4-point improvement in Itch NRS at Week 1 among patients with baseline Itch NRS score ≥4
- Null Hypotheses[ITCH D2]: Proportion of baricitinib 4-mg patients achieving a 4-point improvement in Itch NRS is equal to the proportion of placebo patients achieving a 4-point improvement in Itch NRS at Day 2 among patients with baseline Itch NRS score ≥4
- Null Hypotheses[EASI75]: Proportion of baricitinib 2-mg patients achieving EASI75 is equal to the proportion of placebo patients achieving EASI75 at Week 16
- Null Hypotheses[ITCH W16]: Proportion of baricitinib 2-mg patients achieving a 4-point improvement in Itch NRS is equal to the proportion of placebo patients achieving a 4-point improvement in Itch NRS at Week 16 among patients with baseline Itch NRS score ≥4

- Null Hypotheses[EASI PCFB]: Percent change from baseline in EASI score for baricitinib 2-mg patients is equal to the percent change from baseline in EASI score for placebo patients at Week 16
- Null Hypotheses[ITCH W4]: Proportion of baricitinib 2-mg patients achieving a 4-point improvement in Itch NRS is equal to the proportion of placebo patients achieving a 4-point improvement in Itch NRS at Week 4 among patients with baseline Itch NRS score ≥ 4
- Null Hypotheses[SCORAD75]: Proportion of baricitinib 2-mg patients achieving SCORAD75 is equal to the proportion of placebo patients achieving SCORAD75 at Week 16
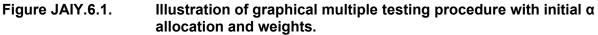- Null Hypotheses[EASI90]: Proportion of baricitinib 2-mg patients achieving EASI90 is equal to the proportion of placebo patients achieving EASI90 at Week 16
- Null Hypotheses[PAIN NRS]: Mean change from baseline in Skin Pain NRS for baricitinib 2-mg patients is equal to the mean change from baseline in Skin Pain NRS for placebo patients at Week 16
- Null Hypotheses[ADSS2 W16]: Mean change from baseline in the score of Item 2 of the ADSS for baricitinib 2-mg patients is equal to the mean change from baseline in the score of Item 2 of the ADSS for placebo patients at Week 16
- Null Hypotheses [ITCH W2]: Proportion of baricitinib 2-mg patients achieving a 4-point improvement in Itch NRS is equal to the proportion of placebo patients achieving a 4-point improvement in Itch NRS at Week 2 among patients with baseline Itch NRS score ≥ 4
- Null Hypotheses[ADSS2 W1]: Mean change from baseline in the score of Item 2 of the ADSS for baricitinib 2-mg patients is equal to the mean change from baseline in the score of Item 2 of the ADSS for placebo patients at Week 1
- Null Hypotheses[ITCH W1]: Proportion of baricitinib 2-mg patients achieving a 4-point improvement in Itch NRS is equal to the proportion of placebo patients achieving a 4-point improvement in Itch NRS at Week 1 among patients with baseline Itch NRS score ≥ 4
- Null Hypotheses[ITCH D2]: Proportion of baricitinib 2-mg patients achieving a 4-point improvement in Itch NRS is equal to the proportion of placebo patients achieving a 4-point improvement in Itch NRS at Day 2 among patients with baseline Itch NRS score ≥ 4

The multiple testing strategy for the primary and the major secondary endpoint will be implemented through the graphical testing procedure depicted by Figure JAIY.6.1.

Figure JAIY.6.1. **Illustration of graphical multiple testing procedure with initial α allocation and weights.**

For Japan, the same testing strategy will be used, however, the proportion of patients achieving EASI75 and IGA of 0 or 1 at Week 16 are considered co-primary. Both endpoints have to meet statistical significance compared with placebo in order to demonstrate a superiority of a given baricitinib dose (baricitinib 4-mg and/or baricitinib 2-mg) compared with placebo.

There will be no adjustment for multiple comparisons for any other analyses.

6.7. Patient Disposition

An overview of patient populations will be summarized by treatment group. Frequency counts and percentages of patients excluded prior to randomization by primary reason for exclusion will be provided for patients who failed to meet study entry requirements during screening.

Patient disposition through Week 16 will be summarized using the ITT population. Frequency counts and percentages of patients who complete the study treatment visits or discontinue early from the study along with whether they completed follow-up, did not complete follow-up or

enrolled into the extension will be summarized separately by treatment group for patients who are not rescued and for patients who are rescued, along with their reason for study discontinuation. Frequency counts and percentages of patients who complete the treatment or discontinue treatment early will also be summarized separately by treatment group for patients who are not rescued and for patients who are rescued, along with their reason for treatment discontinuation.

A listing of patient disposition will be provided for all randomized patients, with the extent of their participation in the study and the reason for discontinuation. A listing of all randomized patients with their treatment assignment will also be provided.

6.8. Patient Characteristics

Patient characteristics including demographics and baseline characteristics will be summarized descriptively by treatment group for the ITT population. Historical illnesses and pre-existing conditions will be summarized descriptively by treatment group for the ITT population. No formal statistical comparisons will be made among treatment groups unless otherwise stated.

6.8.1. Demographics

Patient demographics will be summarized as described above. The following demographic information will be included:

- Age
- Age group (<65 vs. ≥65)
- Age group (<65, ≥65 to <75, ≥75 to <85, ≥85)
- Gender (male, female)
- Race (American Indian or Alaska Native, Asian, Black or African American, Native Hawaiian or Other Pacific Islander, White, Multiple)
- Region (as defined in Table JAIY.5.1)
- Country
- Weight (kg)
- Weight category (<60 kg, ≥60 to <100 kg, ≥100 kg)
- Height (cm)
- BMI (kg/m^2)
- BMI category (<25 kg/m^2, ≥25 to <30 kg/m^2, ≥30 kg/m^2)

A listing of patient demographics will also be provided for the ITT population.

6.8.2. Baseline Disease Characteristics

The following baseline disease information will be categorized and presented for baseline AD clinical characteristics, baseline health outcome measures, and other baseline demographic and disease characteristics as described above:

- Duration since AD diagnosis (years)
- Duration since AD diagnosis category (0 to <2 years, 2 to <5 years, 5 to <10 years, 10 to <20 years, ≥20 years)

- Age at Diagnosis (years)
- Age Group at Diagnosis (<18 years, ≥18 to <50 years, ≥50 years)
- Habits (Alcohol: Never, Current, Former; Tobacco: Never, Current, Former)
- Skin Infections treated with a pharmacological agent within past year (yes, no, unknown; number if yes)
- Atopic Dermatitis Flares within past year (yes, no, unknown; number if yes)
- Validated Investigator's Global Assessment for AD (IGA) score
- Eczema Area and Severity Index (EASI) score
- SCORing Atopic Dermatitis (SCORAD)
- Body Surface Area (BSA) affected by AD
- Hospital Anxiety Depression Scale (HADS) subscales
- Patient-Oriented Eczema Measure (POEM)
- Itch Numerical Rating Scale (NRS)
- Atopic Dermatitis Sleep Scale (ADSS) Item 2
- Dermatology Life Quality Index (DLQI)
- Skin Pain NRS
- Patient Global Impression of Severity (PGI-S-AD)
- Prior therapy (topical therapy only; systemic therapy)
- Prior use of Cyclosporine (yes, no)
- Cyclosporine inadequate response (yes, no)
- Cyclosporine intolerance (yes, no)
- Cyclosporine contraindication [ineligible] (yes, no)
- Cyclosporine inadvisable (yes no)
- Prior use of TCNI (yes, no)
- TCNI inadequate response (yes, no)
- TCNI intolerance (yes, no)
- TCNI contraindication [ineligible] (yes, no)
- TCNI inadvisable (yes, no)
- Vaccine (yes, no)
- Baseline renal function status: impaired (eGFR <60 mL/min/1.73 m^2) or not impaired (eGFR ≥60 mL/min/1.73 m^2)
- Immunoglobulin E (IgE): intrinsic(<200 kU/I) or extrinsic (≥200 kU/I)

6.8.3. *Historical Illness and Pre-existing Conditions*

Historical illnesses are defined as those conditions recorded in the Pre-existing Conditions and Medical History electronic case report form (eCRF) or from the Prespecified Medical History: Comorbidities eCRF with an end date prior to the informed consent date. The number and percentage of patients with selected historical diagnoses will be summarized by treatment group using the ITT population. Historical diagnoses will be categorized using the Medical Dictionary for Regulatory Activities (MedDRA®, most current available version) algorithmic standardized MedDRA queries (SMQs) or similar pre-defined lists of preferred terms (PTs) of interest.

Preexisting conditions are defined as those conditions with a start date prior to the first dose of the study drug and stop dates that are at or after the informed consent date or have no stop date

(i.e., are ongoing). For events occurring on the day of the first dose of study treatment, the date and time of the onset of the event will both be used to determine if the event was pre-existing. Conditions with a partial or missing start date (or time if needed) will be assumed to be 'not pre-existing' unless there is evidence, through comparison of partial dates, to suggest otherwise. Pre-existing conditions will be categorized using the MedDRA SMQs or similar pre-defined lists of PTs of interest. Frequency counts and percentages of patients with selected pre-existing conditions will be summarized by treatment group using the ITT population.

6.9. Treatment Compliance

Patient compliance with study medication will be assessed from Week 0 (Visit 2) to Week 16 (Visit 8) or Early Termination using the ITT population.

All patients are expected to take 2 tablets daily from a package as described in the protocol. Each bottle contains 36 tablets. A patient is considered noncompliant if he or she misses >20% of the prescribed doses during the study, unless the patient's study drug is withheld by the investigator. For patients who had their treatment temporarily interrupted by the investigator, the period of time that dose was withheld will be taken into account in the compliance calculation.

Compliance in the period of interest up to Visit x will be calculated as follows:

$$\text{Compliance} = \frac{\text{total number of tablets dispensed} - \text{total number of tablets returned}}{\text{expected number of total tablets}}$$

where
- Total number of tablets dispensed: sum of tablets dispensed in the period of interest prior to Visit x;
- Total number of tablets returned: sum of the tablets returned in the period of interest prior to and including Visit x;
- Expected number of tablets: number of days in the period of interest*number of tablets taken per day = [(date of last dose – date of first dose + 1) – number of days of temporary drug interruption]*number of tablets taken per day

Patients who are significantly noncompliant (compliance <80%) through Week 16 will be excluded from the PPS population.

Descriptive statistics for percent compliance and non-compliance rate will be summarized for the ITT population by treatment group for Week 0 through Week 16. Sub-intervals of interest, such as compliance between visits, may also be presented. The number of expected doses, tablets dispensed, tablets returned, and percent compliance will be listed by patient for Week 0 through Week 16.

6.10. Rescue Therapy

Rescue therapy with additional topical and systemic therapies is available starting after 2 weeks of treatment (Visit 4), for patients who are experiencing worsening and unacceptable symptoms of AD despite treatment with IP and moderate-potency TCS. Patients whose lesions persist or worsen despite the use of emollients and background TCS therapy (e.g., triamcinolone 0.1% cream and/or hydrocortisone 2.5% ointment) and/or patients who require prolonged applications

of triamcinolone 0.1% cream (moderate-potency TCS) on large surfaces may be considered for rescue to high- or ultra-high-potency TCS. If topical rescue therapy as described above fails to sufficiently control AD symptoms, then oral systemic medications may be used as rescue (e.g., corticosteroids, cyclosporine, methotrexate); however, investigational product will be required to be permanently discontinued for the remainder of the 16-week study duration. If these medications are needed for other medical conditions (e.g., asthma flare), they will still be treated as rescue medications.

The initial rescue therapy will be the first non-missing record before the last dose date from the CRF page *Concomitant Therapy: Rescue Therapy*.

A summary of the initial rescue therapy and the reason for requiring initial rescue will be produced, as well as a summary of the proportion of patients initially rescued at each study visit. A summary of all rescue medications will be provided.

6.11. Previous and Concomitant Therapy

Summaries of previous and concomitant medications will be based on the ITT population.

At screening, previous and current AD treatments are recorded for each patient. Concomitant therapy for the treatment period is defined as therapy that starts before or during the treatment period and ends during the treatment period or is ongoing (has no end date or ends after the treatment period). Should there be insufficient data to make this comparison (for example, the concomitant therapy stop year is the same as the treatment start year, but the concomitant therapy stop month and day are missing), the medication will be considered as concomitant for the treatment period.

Summaries of previous medications will be as follows:

- Previous AD therapies

Summaries of concomitant medications, with sponsor and non-sponsor provided background TCS included, will be as follows:

- General Concomitant medications excluding rescue medicine

6.11.1. Background TCS

Background TCS therapy with moderate-potency and/or low-potency TCS (e.g., triamcinolone 0.1% cream and/or hydrocortisone 2.5% ointment) are to be used on active lesions, as described in Section 7.7.3 in the protocol.

The dispensed weight of sponsor-provided TCS tubes for the two different potencies (low and moderate) varies between countries due to different supply regions. Average weights of full tubes were used to determine the dispensed weights for each region. Returned tubes were weighed with cap (without the carton) to determine the amount of TCS in grams (g) used at each visit.

For low potency TCS, the dispensed tube weight with cap (without the carton) in Japan is 13.5g. For countries supplied by European distributors (Austria, Germany, Italy, Poland, and Spain), the dispensed weight of low potency TCS is 21g. The remaining countries, supplied by US distributors (Argentina, Australia, Korea, and Taiwan), the weight of low potency TCS is 40g.

For moderate potency TCS, the dispensed tube weight with cap (without the carton) in Japan is 13.5g. For countries supplied by European distributors, the dispensed weight of moderate potency TCS is 38g. The remaining countries, supplied by US distributors, the weight of moderate potency TCS is 40g.The total amount of background TCS, provided by sponsor, will be summarized in grams by potency (low and moderate) and both potencies, between visits (Week 0 through Week 1, Week 1 through Week 2, Week 2 through Week 4, Week 4 through Week 8, Week 8 through 12, Week 12 through Week 16), and throughout the entire 16-week treatment period. If a returned tube is not weighed or not returned, then the tube can be classified as partially used, fully used, unused, or unknown. Partially used rescue medication tubes will be defined as 50% used whereas fully used and unused tubes will be defined as 100% and 0% used respectively. When drug accountability is not performed for a particular tube of rescue medication or an answer of 'unknown' is given for a tube which is not returned, that particular tube will not be included in the analysis. The main analysis on the total amount of background TCS throughout the entire 16-week treatment period will apply censoring rule #1. After patients who get rescued or discontinue IP, whichever is earlier, it is assumed that they would use the same amount of TCS as they did before. Analysis will be done via analysis of variance (ANOVA), with geographic region, baseline disease severity and treatment as factors in the model. The secondary analysis will apply censoring rule #2 with the same assumptions as described above.

Whether any background TCS is used or not used for each patient is also collected on the diary device in each day starting from the first dose date throughout the study.

The total number of days that the patients did not use background TCS will be summarized by both potencies throughout the entire 16-week treatment period. The main analysis applies censoring rule #1. After patients who are rescued or discontinue IP, it is assumed that background TCS would be applied each day. In case of missing values in the daily diary, it will be assumed that background TCS has been used. Analysis will be done via ANOVA, with geographic region, baseline disease severity and treatment as factors in the model. A secondary analysis will apply censoring rule #2, with the same assumptions for missing values as described above.

6.12. Efficacy Analyses

The general methods used to summarize efficacy data, including the definition of baseline value for assessments are described in Section 6.2. The censoring rules applied to data as well as imputation methods are described in Section 6.4.

Table JAIY.6.4 provides the descriptions and derivations of the primary, secondary, and exploratory efficacy outcomes.

Table JAIY.6.5 provides the detailed analyses including analysis type, method and imputation, population, time point, and comparisons for efficacy analyses.

Table JAIY.6.4.   **Description and Derivation of Primary, Secondary and Exploratory Efficacy Outcomes**

Measure	Description	Variable	Derivation / Comment	Imputation Approach if Missing Components
Validated Investigator's Global Assessment for AD (IGA)	The validated Investigator's global assessment of the patient's overall severity of their AD, based on a static, numeric 5-point scale from 0 (clear) to 4 (severe). The score is based on an overall assessment of the degree of erythema, papulation/induration, oozing/crusting, and lichenification.	IGA score	Single item. Range: 0 to 4 0 represents "clear" 4 represents "severe"	Single item, missing if missing.
		Change from baseline in IGA score	Change from baseline: observed IGA score – baseline IGA score	Missing if baseline or observed value is missing.
		▪ IGA [0,1] with ≥2-point improvement ▪ IGA [0]	▪ Observed score of 0 or 1 and change from baseline ≤2 ▪ Observed score of 0	▪ Missing if baseline or observed value is missing. ▪ Single item, missing if missing.

Eczema Area and Severity Index (EASI)	The EASI assesses objective physician estimates of 2 dimensions of atopic dermatitis – disease extent and clinical signs (Hanifin et al 2001) – by scoring the extent of disease (percentage of skin affected: 0 = 0%; 1 = 1-9%; 2 = 10-29%; 3 = 30-49%; 4 = 50-69%; 5 = 70-89%; 6 = 90-100%) and the severity of 4 clinical signs (erythema, edema/papulation, excoriation, and lichenification) each on a scale of 0 to 3 (0 = none, absent; 1 = mild; 2 = moderate; 3 = severe) at 4 body sites (head and neck, trunk, upper limbs, and lower limbs). Half scores are allowed between severities 1, 2 and 3. Each body site will have a score that ranges from 0 to 72, and the final EASI score will be obtained by weight-averaging these 4 scores. Hence, the final EASI score will range from 0 to 72 for each time point.	EASI score	Derive EASI region score for each of head and neck, trunk, upper limbs, and lower limbs as follows: $EASI_{region}$ = (Erythema + edema/papulation + Excoriation + Lichenification) *(value from percentage involvement), where erythema, edema/papulation, excoriation, and lichenification are evaluated on a scale of 0 to 3 and value from percentage involvement is on a scale of 0 to 6. Then total EASI score is as follows: $EASI = 0.1*EASI_{head\ and\ neck} + 0.3*EASI_{trunk} + 0.2*EASI_{upper\ limbs} + 0.4*EASI_{lower\ limbs}$	N/A – partial assessments cannot be saved.
		▪ Change from baseline in EASI score ▪ Percent change from baseline EASI score	Change from baseline: observed EASI score – baseline EASI score % change from baseline: $100 \times \dfrac{Observed\ score - Baseline}{Baseline}$	Missing if baseline or observed value is missing.
		EASI50	% Improvement in EASI score from baseline ≥ 50%: % change from baseline ≤ -50	Missing if baseline or observed value is missing.
		EASI75	% Improvement in EASI score from baseline ≥75%: % change from baseline ≤-75	Missing if baseline or observed value is missing.
		EASI90	% Improvement in EASI score from baseline ≥90%: % change from baseline ≤-90	Missing if baseline or observed value is missing.
Body Surface Area (BSA) Affected by AD	Body surface area affected by AD will be assessed for 4 separate body regions and is collected as part of the EASI assessment: head and neck, trunk (including genital region), upper extremities, and lower extremities	BSA score	Use the percentage of skin affected for each region (0 to 100%) in EASI as follows: $BSA\ Total = 0.1*BSA_{head\ and\ neck} + 0.3*BSA_{trunk} + 0.2*\ BSA_{upper\ limbs} +$	N/A – partial assessments cannot be saved.

			0.4*BSA$_{lower\ limbs}$	
	(including the buttocks). Each body region will be assessed for disease extent ranging from 0% to 100% involvement. The overall total percentage will be reported based off of all 4 body regions combined, after applying specific multipliers to the different body regions to account for the percent of the total BSA represented by each of the 4 regions.	Change from baseline in BSA score	Change from baseline: observed BSA score – baseline BSA score	Missing if baseline or observed value is missing.
SCORing Atopic Dermatitis (SCORAD)	The SCORing Atopic Dermatitis (SCORAD) index uses the rule of nines to assess disease extent (head and neck 9%; upper limbs 9% each; lower limbs 18% each; anterior trunk 18%; back 18%; and genitals 1%). It evaluates 6 clinical characteristics to determine disease severity: (1) erythema, (2) edema/papulation, (3) oozing/crusts, (4) excoriation, (5) lichenification, and (6) dryness on a scale of 0 to 3 (0=absence, 1=mild, 2=moderate, 3=severe). The SCORAD index also assesses subjective symptoms of pruritus and sleep loss in the last 72 hours on visual analogue scales (VAS) of 0 to 10 where 0 is no itch or sleep loss and 10 is worst imaginable itch or sleep loss. These 3 aspects: extent of disease, disease severity, and subjective symptoms combine to give a maximum possible score of 103 (Stalder et al. 1993; Kunz et al. 1997; Schram et al. 2012).	SCORAD score	SCORAD = A/5 + 7B/2 + C, where A is extent of disease, range 0-100 B is disease severity, range 0-18 C is subjective symptoms, range 0-20	Missing if components A and B are missing or if component C is missing. Partial assessments performed by physician cannot be saved and partial assessments performed by subject cannot be saved.
		• Change from baseline in SCORAD score • Percent change from baseline in SCORAD score	Change from baseline: observed SCORAD score – baseline SCORAD score % change from baseline: $100 \times \dfrac{Observed\ score - Baseline}{Baseline}$	Missing if baseline or observed value is missing.
		SCORAD75	% Improvement in SCORAD from baseline ≥75%: % change from baseline ≤-75	Missing if baseline or observed value is missing.
		SCORAD90	% Improvement in SCORAD from baseline ≥90%: % change from baseline ≤-90	Missing if baseline or observed value is missing.

Table JAIY.6.5. **Description of Primary, Secondary and Exploratory Efficacy Analyses**

Measure	Variable	Analysis Method (Section 6.2.3)	Population (Section 6.2.1)	Comparison/Time Point	Analysis Type
Validated Investigator's Global Assessment for AD (IGA)	Proportion of patients achieving IGA [0,1] with a ≥2-point improvement	Logistic regression using NRI	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Primary analysis
			PPS	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Sensitivity analysis
			ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 4	Secondary analysis
		Logistic regression using pMI and Tipping Point	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Sensitivity analysis
	Proportion of patients achieving IGA [0]	Logistic regression using NRI	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Secondary analysis
Eczema Area and Severity Index (EASI)	• EASI score • Change from baseline in EASI score • Percent change from baseline in EASI score	MMRM	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Key secondary analysis
			PPS	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Sensitivity analysis
		ANCOVA using mLOCF	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Sensitivity analysis
		pMI and Tipping Point	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Sensitivity analysis
	• Proportion of patients achieving EASI50	Logistic regression using NRI	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Secondary analysis
	• Proportion of patients achieving EASI75 • Proportion of patients achieving EASI90	Logistic regression using NRI	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Key secondary analysis
			PPS	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Sensitivity analysis
		pMI and Tipping Point	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Sensitivity analysis
Body Surface Area (BSA) Affected by AD	• BSA score • Change from baseline in BSA score	MMRM	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Secondary analysis
		ANCOVA using mLOCF	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Sensitivity analysis

Measure	Variable	Analysis Method (Section 6.2.3)	Population (Section 6.2.1)	Comparison/Time Point	Analysis Type
SCORing Atopic Dermatitis (SCORAD)	• SCORAD score • Change from baseline in SCORAD score • Percent change from baseline in SCORAD score	MMRM	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Secondary analysis
		ANCOVA using mLOCF	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Sensitivity analysis
	Proportion of patients achieving SCORAD75	Logistic regression using NRI	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Key secondary analysis
			PPS	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Sensitivity analysis
		Logistic regression using pMI	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Sensitivity analysis
	Proportion of patients achieving SCORAD90	Logistic regression using NRI	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Secondary analysis
		pMI and Tipping Point	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Sensitivity analysis
Skin Infections	Proportion of patients developing skin infections requiring antibiotic treatment	Fisher's exact	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Secondary analysis

Abbreviations: ANCOVA = analysis of covariance; Bari = baricitinib; ITT = intent-to-treat; mLOCF = modified last observation carried forward;

MMRM = mixed model repeated measures; NRI = nonresponder imputation; PBO = placebo; pMI=placebo multiple imputation; PPS = per protocol set.

Notes: (1) for all other post-baseline visits not mentioned in the table, but collected for the measures as specified in the protocol, the analyses will be made as exploratory analyses.

(2) All primary and key secondary analyses will be performed for Japan population. Other key secondary and exploratory analysis may be performed for Japan population.

6.12.1. Primary Outcome and Methodology

The validated Investigator's Global Assessment for AD (IGA) uses the clinical characteristics of erythema, papulation/induration, oozing/crusting and lichenification to produce a single-item score ranging from 0 to 4. The primary analysis of the study is to test the null hypotheses that neither baricitinib 4 mg nor baricitinib 2 mg is superior to placebo when evaluating the proportion of patients achieving IGA of 0 or 1 at Week 16 in the ITT population. The analysis assumes that treatment response disappears after patients are rescued or permanently discontinue from treatment. This will serve as the primary estimand. In this estimand, missing data due to the application of the primary censoring rule and the occurrence of other non-censor intercurrent events will be imputed using the NRI method described in Section 6.4.1.

A supplemental estimand is to test the null hypotheses that neither baricitinib 4 mg nor baricitinib 2 mg is superior to placebo when evaluating the proportion of patients achieving IGA of 0 or 1 at Week 16 in the ITT population. This analysis assumes the treatment response disappears after patients permanently discontinue from treatment. In this supplemental estimand, missing data due to the application of the secondary censoring rule and the occurrence of other non-censor intercurrent events will be imputed using the NRI method described in Section 6.4.1.

A logistic regression analysis as described in Section 6.2.3 will be used for the comparisons. The odds ratio, the corresponding 95% CIs and p-value, as well as the treatment differences and the corresponding 95% CIs, will be reported.

Multiplicity controlled analyses will be performed on the primary and key secondary (see Section 4.2.1) objectives to control the overall Type I error rate at a 2-sided alpha level of 0.05. A graphical approach will be used to perform the multiplicity controlled analyses as described in Section 6.6.

6.12.2. Secondary and Exploratory Efficacy Analyses

For secondary analysis, the null hypotheses is that neither baricitinib 4 mg nor baricitinib 2 mg is superior to placebo in the ITT population. These analyses assume treatment response disappears after patients are rescued or permanently discontinued from treatment and will serve as the primary estimand. In this estimand, missing data due to the application of the primary censoring rule and the occurrence of other non-censor intercurrent events will be imputed using the method described in Table JAIY.6.1.

A supplemental estimand for secondary endpoints is to test the null hypotheses that neither baricitinib 4 mg nor baricitinib 2 mg is superior to placebo in the ITT population. These analyses assume the treatment response disappears after patients permanently discontinue from treatment. In this supplemental estimand, missing data due to the application of the secondary censoring rule and the occurrence of other non-censor intercurrent events will be imputed using the method described in Table JAIY.6.1.

A list of exploratory endpoints are provided in Section 4.2.2. There will be no adjustment for multiple comparisons for exploratory endpoints. The secondary and exploratory efficacy

endpoints are detailed in Table JAIY.6.4 and analyses are provided in Table JAIY.6.5. Health outcomes analyses are described in Section 6.13.

6.12.3. Sensitivity Analyses

Sensitivity analyses are included to demonstrate robustness of analyses methods using different missing data imputations, censoring rules, populations and analyses assumptions. Sensitivity analyses for select outcomes have been previously described and include the following:

- Analyses of key endpoints using the per-protocol analysis set (Section 6.2.1)
- Analyses of key endpoints using the secondary censoring rule (Section 6.2)
- Placebo multiple imputation (Section 6.4.4)
- Tipping point analysis (Section 6.4.5)
- The addition of a treatment-by-region interaction to the logistic regression model for the primary outcome (Section 6.5)
- Analysis of continuous outcomes with ANCOVA (Section 6.2.3), with missing data imputed using mLOCF (Section 6.4.3).

6.13. Health Outcomes/Quality-of-Life Analyses

The general methods used to summarize health outcomes and quality-of-life measures, including the definition of baseline value for assessments are described in Section 6.1.

Health outcomes and quality-of-life measures will generally be analyzed according to the formats discussed in Section 6.12.

Table JAIY.6.6 includes the descriptions and derivations of the health outcomes and quality-of-life measures.

Table JAIY.6.7 provides the detailed analyses including analysis type, method and imputation, population, time point, and comparisons for health outcomes and quality-of-life measures.

Table JAIY.6.6.   **Description and Derivation of Health Outcomes and Quality-of-Life Measures**

Measure	Description	Variable	Derivation / Comment	Imputation Approach if Missing Components
Itch Numeric Rating Scale (NRS)	The Itch Numeric Rating Scale (NRS) is a patient-administered, 11-point horizontal scale anchored at 0 and 10, with 0 representing "no itch" and 10 representing "worst itch imaginable." Overall severity of a patient's itching is indicated by selecting the number that best describes the worst level of itching in the past 24 hours (Naegeli et al. 2015; Kimball et al. 2016). Refer to Section 6.2.2 for details on how to calculate the weekly score which will be used in the continuous analysis.	Itch NRS score	Single item; range 0-10. Refer to Section 6.2.2 on how to derive the visit score.	Refer to Section 6.2.2 on how to derive the weekly visit score.
		• Change from baseline in Itch NRS • Percent change from baseline in Itch NRS	Change from baseline: observed Itch score – baseline Itch score % change from baseline: $100 \times \dfrac{Observed\ score - Baseline}{Baseline}$	Missing if baseline or observed value is missing.
		4-point Itch improvement in subgroup of patients with baseline Itch NRS ≥4	Change from baseline ≤-4 and baseline ≥4	Missing if baseline is missing or <4 or observed value is missing.
		Itch-free days (Itch NRS = 0)	The number of itch-free days during intervals starting on the day of the first study drug administration. This will be calculated for the following intervals: baseline to Week 4, Week 4 to Week 8, Week 8 to Week 12 and Week 12 to Week 16. Day 1 is defined as the day of first study drug administration therefore the baseline to Week 4 assessment is based on Day 1 to Day 28, Week 4 to Week 8 is based on Day 29 to Day 56, etc.	Missing if observed value is missing.

Measure	Description	Variable	Derivation / Comment	Imputation Approach if Missing Components
Skin Pain Numeric Rating Scale (NRS)	Skin Pain NRS is a patient-administered, 11-point horizontal scale anchored at 0 and 10, with 0 representing "no pain" and 10 representing "worst pain imaginable." Overall severity of a patient's skin pain is indicated by selecting the number that best describes the worst level of skin pain in the past 24 hours Refer to Section 6.2.2 for details on how to calculate the weekly score which will be used in the continuous analysis.	Skin Pain NRS score	Single item; range 0 to 10. Refer to Section 6.2.2 on how to derive the visit score.	Refer to Section 6.2.2 on how to derive the visit score.
		Change from baseline in Skin Pain NRS	Change from baseline: observed skin pain score – baseline skin pain score	Missing if baseline or observed value is missing.
		Skin Pain-free days (Skin Pain NRS = 0)	The number of skin pain-free days during intervals starting on the day of the first study drug administration. This will be calculated for the following intervals: baseline to Week 4, Week 4 to Week 8, Week 8 to Week 12 and Week 12 to Week 16. Thus, if Day 1 is defined as the day of first study drug administration, the baseline to Week 4 assessment is based on Day 1 to Day 28, Week 4 to Week 8 is based on Day 29 to Day 56, etc.	Missing if observed value is missing.

Measure	Description	Variable	Derivation / Comment	Imputation Approach if Missing Components
Atopic Dermatitis Sleep Scale (ADSS)	The Atopic Dermatitis Sleep Scale (ADSS) is a 3-item, patient-administered questionnaire developed to assess the impact of itch on sleep including difficulty falling asleep, frequency of waking, and difficulty getting back to sleep last night. Patient's rate their difficulty falling asleep and difficulty getting back to sleep, items 1 and 3, respectively, using a 5-point Likert-type scale with response options ranging from 0 "not at all" to 4 "very difficult." Patients report their frequency of waking last night, item 2, by selecting the number of times they woke up each night, ranging from 0 to 29 times. The ADSS is designed to be completed each day with respondents thinking about sleep "last night." Each item is scored individually.	▪ Item 1 score of ADSS ▪ Item 2 score of ADSS ▪ Item 3 score of ADSS	Single items: Item 1, range 0 to 4; Item 2, range 0 to 29; Item 3, range 0 to 4. Refer to Section 6.2.2 on how to derive the visit score.	Refer to Section 6.2.2 on how to derive the weekly visit score.
		▪ Change from baseline in score of Item 1 of ADSS ▪ Change from baseline in score of Item 2 of ADSS ▪ Change from baseline in score of Item 3 of ADSS	Change from baseline: observed ADSS item score – baseline ADSS item score	Missing if baseline or observed value is missing.

Measure	Description	Variable	Derivation / Comment	Imputation Approach if Missing Components
Patient- Oriented Eczema Measure (POEM)	The POEM is a simple, 7-item, patient-administered scale that assesses disease severity in children and adults. Patients respond to questions about the frequency of 7 symptoms (itching, sleep disturbance, bleeding, weeping/oozing, cracking, flaking, and dryness/roughness) over the last week. Response categories include "No days," "1-2 days," "3-4 days," "5-6 days," and "Every day" with corresponding scores of 0, 1, 2, 3, and 4, respectively. Scores range from 0-28 with higher total scores indicating greater disease severity (Charman et al. 2004).	POEM score	POEM total score: sum of questions 1 to 7, Range 0 to 28.	If a single question is left unanswered, then that question is scored as 0. If more than one question is unanswered, then the tool is not scored. If more than one response is selected, then the response with the highest score is used.
		Change from baseline in POEM score	Change from baseline: observed POEM score – baseline POEM score	Missing if baseline or observed value is missing.
		4-point improvement in POEM score in subgroup of patients with baseline ≥4	Change from baseline ≤-4 and baseline ≥4	Missing if baseline is missing or <4 or observed value is missing.

Measure	Description	Variable	Derivation / Comment	Imputation Approach if Missing Components
Patient Global Impression of Severity–Atopic Dermatitis (PGI-S-AD)	The Patient Global Impression of Severity–Atopic Dermatitis (PGI-S-AD) is a single-item question asking the patient how they would rate their overall AD symptoms over the past 24 hours. The 5 categories of responses range from "no symptoms" to "severe."	PGI-S-AD score	Single item. Range 1 to 5. Refer to Section 6.2.2 on how to derive the visit score.	Refer to Section 6.2.2 on how to derive the visit score.
		Change from baseline in PGI-S-AD	Change from baseline: observed PGI-S-AD score – baseline PGI-S-AD score	Missing if baseline or observed value is missing.
Hospital Anxiety Depression Scale (HADS)	The Hospital Anxiety Depression Scale (HADS) is a 14-item self-assessment scale that determines the levels of anxiety and depression that a patient is experiencing over the past week. The HADS utilizes a 4-point Likert scale (e.g., 0 to 3) for each question and is intended for ages 12 to 65 years (Zigmond and Snaith 1983; White et al. 1999). Scores for each domain (anxiety and depression) can range from 0 to 21, with higher scores indicating greater anxiety or depression (Zigmond and Snaith 1983; Snaith 2003).	HADS score for anxiety and depression domains	Anxiety domain score is sum of the seven anxiety questions, range 0 to 21; Depression domain score is sum of the seven depression questions, range 0 to 21.	N/A – partial assessments cannot be saved.
		Change from baseline in HADS total score, anxiety and depression domain	Change from baseline: observed HADS domain score – baseline HADS domain score	Missing if baseline or observed value is missing.
Dermatology Life Quality Index (DLQI)	The Dermatology Life Quality Index (DLQI) is a simple, patient-administered, 10-item, validated, quality-of-life questionnaire that covers 6 domains including symptoms and feelings, daily activities, leisure, work and school, personal relationships, and treatment. The recall period of this scale is over the "last week." Response categories include "a little," "a lot," and "very much," with corresponding	Symptoms and feelings domain	Sum of questions 1 and 2, range 0 to 6.	N/A – partial assessments cannot be saved.
		Daily activities domain	Sum of questions 3 and 4, range 0 to 6.	N/A – partial assessments cannot be saved.
		Leisure domain	Sum of questions 5 and 6, range 0 to 6.	N/A – partial assessments cannot be saved.

Measure	Description	Variable	Derivation / Comment	Imputation Approach if Missing Components
	scores of 1, 2, and 3, respectively, and "not at all," or unanswered ("not relevant") responses scored as 0. Scores range from 0-30 with higher scores indicating greater impairment of quality of life. A DLQI total score of 0 to 1 is considered as having no effect on a patient's health-related QoL (Hongbo et al. 2005), and a 4-point change from baseline is considered as the minimal clinically important difference threshold (Khilji et al. 2002; Basra et al. 2015).	Work and school domain	Sum of questions 7 and 7B (if it is answered), range 0 to 3. Responses of "yes" and "no" on Question 7 are given scores of 3 and 0 respectively. If Question 7 is answered "no" then Question 7b is answered with "a lot", "a little", "not at all" getting scores of 2, 1, 0 respectively.	N/A – partial assessments cannot be saved.
		Personal relationships domain	Sum of questions 8 and 9, range 0 to 6.	N/A – partial assessments cannot be saved.
		Treatment domain	Question 10, range 0 to 3.	N/A – partial assessments cannot be saved.
		DLQI total score	DLQI total score: sum of all six DLQI domain scores, range 0 to 30.	N/A – partial assessments cannot be saved.
		Change from baseline in DLQI	Change from baseline: observed DLQI score – baseline DLQI score	Missing if baseline or observed value is missing.
		4-point improvement in DLQI total score in subgroup of patients with baseline ≥4	Change from baseline ≤4 and baseline ≥4	Missing if baseline is missing or <4 or observed value is missing.
		DLQI (0,1)	A DLQI (0,1) response is defined as a post-baseline DLQI total score of 0 or 1	Missing if the DLQI total score is missing

Measure	Description	Variable	Derivation / Comment	Imputation Approach if Missing Components
Work Productivity and Activity Impairment: Atopic Dermatitis (WPAI-AD)	The Work Productivity and Activity Impairment Questionnaire–Atopic Dermatitis (WPAI-AD) records impairment due to AD during the past 7 days. The WPAI-AD consists of 6 items grouped into 4 domains: absenteeism (work time missed), presenteeism (impairment at work/reduced on-the-job effectiveness), work productivity loss (overall work impairment/absenteeism plus presenteeism), and activity impairment. Scores are calculated as impairment percentages (Reilly et al. 1993), with higher scores indicating greater impairment and less productivity.	Employment status	Question (Q)1	Single item, missing if missing.
		Change in employment status	Employed at baseline and remained employed: Q1 = 1 at post-baseline visit and at baseline visit. Not employed at baseline and remain unemployed: Q1 = 0 at post-baseline visit and at baseline visit.	Missing if baseline or observed value is missing.
		Percentage of absenteeism	Percent work time missed due to problem: $(Q2/(Q2 + Q4))*100$	If Q2 or Q4 is missing, then missing.
		Change from baseline in absenteeism	Change from baseline: observed absenteeism – baseline absenteeism	Missing if baseline or observed value is missing.
		Percentage of presenteeism	Percent impairment (reduced productivity while at work) while working due to problem: $(Q5/10)*100$	If Q5 is missing, then missing.
		Change from baseline in presenteeism	Change from baseline: observed presenteeism – baseline absenteeism	Missing if baseline or observed value is missing.
		Overall work impairment	Percent overall work impairment (combines absenteeism and presenteeism) due to problem: $(Q2/(Q2+Q4) + [(1-Q2/(Q2+Q4))*(Q5/10)])*100$	If Q2, Q4, or Q5 is missing, then missing.

LY3009104

Measure	Description	Variable	Derivation / Comment	Imputation Approach if Missing Components
		Change from baseline in work impairment	Change from baseline: observed work impairment – baseline work impairment	Missing if baseline or observed value is missing.
		Percentage of impairment in activities	Percent activity impairment (performed outside of work) due to problem: (Q6/10)*100	If Q6 is missing, then missing.
		Change from baseline in impairment in activities	Change from baseline: observed impairment in activities – baseline impairment in activities	Missing if baseline or observed value is missing.
European Quality of Life–5 Dimensions–5 Levels (EQ-5D-5L)	The European Quality of Life–5 Dimensions–5 Levels (EQ-5D-5L) is a standardized measure of health status that provides a simple, generic measure of health for clinical and economic appraisal. The EQ-5D-5L consists of 2 components: a descriptive system of the respondent's health and a rating of his or her current health state using a 0 to 100 mm VAS. The descriptive system comprises the following 5 dimensions: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. Each dimension has 5 levels: no problems, slight problems, moderate problems, severe problems, and extreme problems. The respondent is asked to indicate his or her health state by ticking (or placing a cross) in the box associated with the most appropriate statement in each of the 5 dimensions. It should be noted that the numerals 1 to 5 have no arithmetic properties and should not be used as an ordinal score. The VAS records the respondent's self-rated health on a vertical VAS where the endpoints are	- EQ-5D mobility - EQ-5D self-care - EQ-5D usual activities - EQ-5D pain/ discomfort - EQ-5D anxiety/ depression	Five health profile dimensions, each dimension has 5 levels: 1 = no problems 2 = slight problems 3 = moderate problems 4 = severe problems 5 = extreme problems It should be noted that the numerals 1 to 5 have no arithmetic properties and should not be used as a primary score.	Each dimension is a single item, missing if missing.
		EQ-5D VAS	Single item. Range 0 to 100. 0 represents "worst health you can imagine" 100 represents "best health you can imagine"	Single item, missing if missing.
		Change from baseline in EQ-5D VAS	Change from baseline: observed EQ-5D VAS score – baseline EQ-5D VAS score	Missing if baseline or observed value is missing.
		EQ-5D-5L UK Population-based index score (health	Derive EQ-5D-5L UK Population-based index score according to the link by using the	N/A – partial assessments cannot be saved

Measure	Description	Variable	Derivation / Comment	Imputation Approach if Missing Components
	labeled "best imaginable health state" and "worst imaginable health state." This information can be used as a quantitative measure of health outcome. The EQ-5D-5L health states, defined by the EQ-5D-5L descriptive system, may be converted into a single summary index by applying a formula that essentially attaches values (also called weights) to each of the levels in each dimension (Herdman et al. 2011; EuroQol Group 2015 [WWW]).	state index)	UK algorithm to produce a patient-level index score between -0.59 and 1.0 (continuous variable).	on the eCOA tablet.
		Change from baseline in EQ-5D-5L UK Population-based index score	Change from baseline: observed EQ-5D-5L UK score – baseline EQ-5D-5L UK score	Missing if baseline or observed value is missing.
		EQ-5D-5L US Population-based index score (health state index)	Derive EQ-5D-5L US Population-based index score according to the link by using the US algorithm to produce a patient-level index score between -0.11 and 1.0 (continuous variable).	N/A – partial assessments cannot be saved on the eCOA tablet.
		Change from baseline in EQ-5D-5L US Population-based index score	Change from baseline: observed EQ-5D-5L US score – baseline EQ-5D-5L US score	Missing if baseline or observed value is missing.
PROMIS Itch Questionnaire (PIQ)	PIQ – Itch Interference: consists of 8 items assessing the impact of itch on various aspects of life. PIQ – Activity and Clothing: consists of 8 items assessing activity and clothing related quality of life impairment from itch in adults "in the past 7 days". PIQ – Mood and Sleep: consists of 8 items assessing mood and sleep related quality of life impairment from itch and impact of itch "in the past 7 days". PIQ – Scratching Behavior: consists of 5 items assessing quality of life impairment from scratching behavior and the physical	PIQ – Itch Interference[a]	8 items. Each range 1 to 5. 1=Never, 2=Rarely, 3=Sometimes, 4=Often, 5=Almost always (continuous variable). Total raw scores are converted to T-Scores with higher scores representing greater impact because of itch.	Calculation is made by HealthMeasures Scoring Service, powered by Assessment Center[SM]
		Change from baseline in PIQ – Itch Interference	Change from baseline: observed PIQ Itch Interference score – baseline PIQ Itch Interference score	Missing if Baseline or observed value is missing.
		PIQ – Activity and Clothing	8 items. Each range 1 to 5. 1=Never, 2=Rarely,	Score is calculated by

Measure	Description	Variable	Derivation / Comment	Imputation Approach if Missing Components
	manifestations of itch in adults "in the past 7 days".		3=Sometimes, 4=Often, 5=Almost always (continuous variable). Total raw scores are converted to T-Scores with higher scores representing greater impact because of itch.	HealthMeasures Scoring Service, powered by Assessment Center^{SM}
		Change from baseline in PIQ – Activity and Clothing	Change from baseline: observed PIQ Activity and Clothing score – baseline PIQ Activity and Clothing score	Missing if Baseline or observed value is missing.
		PIQ – Mood and Sleep	8 items. Each range 1 to 5. 1=Never, 2=Rarely, 3=Sometimes, 4=Often, 5=Almost always (continuous variable). Total raw scores are converted to T-Scores with higher scores representing greater impact because of itch.	Score is calculated by HealthMeasures Scoring Service, powered by Assessment Center^{SM}
		Change from baseline in PIQ – Mood and Sleep	Change from baseline: observed PIQ Mood and Sleep score – baseline PIQ Mood and Sleep score	Missing if Baseline or observed value is missing.
		PIQ – Scratching Behavior	5 items. Each range 1 to 5. Response options for the frequency of scratching behaviors: 1=Never, 2=Rarely, 3=Sometimes, 4=Often, 5=Almost always. The response options for the worry related to scratching items: 1=Not at all, 2=A little bit, 3=Somewhat, 4=Quite a bit, 5=Very much (continuous variable). Total raw scores are converted to T-Scores with higher	Score is calculated by HealthMeasures Scoring Service, powered by Assessment Center^{SM}

Measure	Description	Variable	Derivation / Comment	Imputation Approach if Missing Components
			scores representing more scratching behavior.	
		Change from baseline in PIQ – Scratching Behavior	Change from baseline: observed PIQ Scratching Behavior score – baseline PIQ Scratching Behavior score	Missing if Baseline or observed value is missing.
PROMIS – Sleep Related Impairment	The Sleep Related Impairment Short Form within the PROMIS bank consists of 8 items measuring self-reported perceptions of alertness, sleepiness, and tiredness during usual waking hours, and the perceived functional impairments during wakefulness associated with sleep problems or impaired alertness "in the past 7 days". Response options range from 1=Not at all; 2=A little bit; 3=Somewhat; 4=Quite a bit; to 5=Very much.	PROMIS sleep related impairment	8 items. Each range 1 to 5. 1=Not at all; 2=A little bit; 3=Somewhat; 4=Quite a bit; to 5=Very much (continuous variable). Total raw scores are converted to T-Scores with higher scores representing greater sleep impairment.	Score is calculated by HealthMeasures Scoring Service, powered by Assessment Center[SM]
		Change from baseline in PROMIS sleep related impairment	Change from baseline: observed PROMIS Sleep Related Impairment score – baseline PROMIS Sleep Related Impairment score	Missing if Baseline or observed value is missing.
Neuro-QoL – Cognitive Function	The Cognitive Function Short Form domain within Neuro-QoL bank consists of 8-items measuring Executive Function (perceived difficulties in applications of mental health function related to planning, organizing, calculating, remembering and learning) "in the past 7 days" and General Concerns (perceived difficulties in everyday cognitive abilities such as memory, attention, and decision making) using the lead-in phrase "how much difficulty do you currently have…". The response options for the Executive Function items range from 1=Very Often (several times a day); 2=Often (once a day); 3=Sometimes (2-3 times); 4=Rarely (once); to	Neuro-QoL – Cognitive Function	The total raw scores are converted to T-Scores with higher scores indicating better (desirable) self-reported health.	Score is calculated by HealthMeasures Scoring Service, powered by Assessment Center[SM]
		Change from baseline in Neuro-QoL – Cognitive Function	Change from baseline: observed Neuro-QoL – Cognitive Function score – baseline Neuro-QoL – Cognitive Function score	Missing if Baseline or observed value is missing.

Measure	Description	Variable	Derivation / Comment	Imputation Approach if Missing Components
	5=Never). The response options for the General Concerns items range from 1=Cannot do; 2=A lot; 3=Somewhat; 4=A little, to 5=None.			
Patient Benefit Index (PBI)	The Patient Benefit Index (PBI) measures patient-defined treatment objectives and benefits. It consists of 2 questionnaires. Before therapy, patients complete the standardized "Patient Needs Questionnaire" (PNQ) indicating individual importance of treatment objectives. This reflects their personal preferences with respect to therapeutic benefit. During the study patients rate the extent to which the treatment objectives have been achieved in the "Patient Benefit Questionnaire" (PBQ). Subscales of the Patient Benefit Index are: • Reducing social impairments subscale score: item 11, 13, 14, 15, 16, 17 • Reducing psychological impairments subscale score: item 6, 7, 9, 10, 12 • Reducing impairments due to therapy subscale score: item 18, 19, 20, 21 • Reducing physical impairments subscale score: item 1, 2, 3, 4, 5 • Having confidence in healing subscale score: item 8, 22, 23	• Patient Benefit Index (PBI) global score • Subscale scores	25 items. 0=not at all, 1=somewhat; 2=moderately; 3=quite; 4=very; PBI global score is calculated for each patient by weighing the achievement values of the treatment objectives by their importance to the individual patient. $$PBI = \sum_{i=1}^{k} \frac{PNQ_i}{\sum_{i=1}^{k} PNQ_i} PBQ_i$$ For score calculation, both "does/did not apply" and question unanswered will be treated as missing values. The global score is calculated using only these item pairs, for which the patient has given a response other than "does/did not apply to me" in both PNQ and PBQ. Subscale scores are calculated in the same manner as the global score.	PBI global score and subscales may only be computed if the patient has provided at least 75% valid data in each of the PNQ and PBQ respectively. In this context, the responses "not at all" and "does/did not apply" count as valid data. Thus a treatment goal is regarded missing if the patient has not responded to the item in the PNQ and/or in the PBI.

Measure	Description	Variable	Derivation / Comment	Imputation Approach if Missing Components
		• PBI global score ≥1	PBI global score at least 1	Missing if observed value is missing

^a PIQ – Itch Interference here is the same as PIQ – General in the protocol.

Table JAIY.6.7.　　　Description of Health Outcomes and Quality-of-Life Measures Analyses

Measure	Variable	Analysis Method (Section 6.2.3)	Population (Section 6.2.1)	Comparison/Time Point	Analysis Type
Itch Numeric Rating Scale (NRS)	• Itch NRS score • Change from baseline in Itch NRS score	MMRM	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 1, …, 16	Secondary Analysis
		ANCOVA using mLOCF	ITT	Bari 4 mg or Bari 2 mg vs PBO; Day 2, Week 1,…, 16	Secondary Analysis at Day 2; Sensitivity Analysis for others
		pMI	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 1, …, 16	Sensitivity Analysis
	Percent change from baseline Itch score	MMRM	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 1, …,16	Secondary Analysis
		ANCOVA using mLOCF	ITT	Bari 4 mg or Bari 2 mg vs PBO; Day 2, Week 1, …, 16	Secondary Analysis at Day 2; Sensitivity Analysis for others
	Proportion of patients achieving a 4-point improvement in Itch NRS in subgroup of patients who had baseline Itch NRS ≥4	Logistic regression using NRI	ITT	Bari 4 mg or Bari 2 mg vs PBO; Day 2, Week 1,…, 16	Key Secondary Analysis at Day 2, Week 1, 2 and 16
			PPS	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Sensitivity analysis
		Logistic regression using pMI and Tipping Point	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Sensitivity analysis
	Number of Itch-free (Itch NRS = 0) Days	Two-sample t-test	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 0 to Week 4, …, Week 12 to 16	Exploratory Analysis
Skin Pain Numeric Rating Scale (NRS)	• Skin Pain NRS score • Change from baseline in Skin Pain NRS score	MMRM	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 1,…, 16	Key Secondary Analysis at Week 16
			PPS	Bari 4 mg or Bari 2 mg vs PBO; Week 1,…, 16	Sensitivity analysis

Measure	Variable	Analysis Method (Section 6.2.3)	Population (Section 6.2.1)	Comparison/Time Point	Analysis Type
		ANCOVA using mLOCF	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 1,…, 16	Sensitivity Analysis
		pMI	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 1,…, 16	Sensitivity analysis
	Number of Skin Pain-free (Skin pain NRS = 0) Days	Two-sample t-test	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 0 to Week 4,…, Week 12 to 16	Exploratory Analysis
	Proportion of patients achieving a 4-point improvement in Skin Pain NRS in subgroup of patients with baseline Skin Pain NRS ≥4	Logistic regression using NRI	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 1,…, 16	Other Secondary Analysis
Atopic Dermatitis Sleep Scale (ADSS)	• ADSS item scores • Change from baseline in ADSS item scores	MMRM	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 1,…, 16	Key Secondary Analysis at Week 1 and 16
			PPS	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Sensitivity analysis
		ANCOVA using mLOCF	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 1,…, 16	Sensitivity Analysis
		pMI	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Sensitivity analysis
Patient-Oriented Eczema Measure (POEM)	• POEM score • Change from baseline in POEM score	MMRM	ITT	Bari 4 mg or Bari 2 mg vs PBO; at each post-baseline visit	Secondary Analysis
		ANCOVA using mLOCF	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Sensitivity Analysis
	Proportion of patients achieving a 4-point improvement in subgroup of patients with baseline POEM ≥ 4	Logistic regression using NRI	ITT	Bari 4 mg or Bari 2 mg vs PBO; at each post-baseline visit	Secondary Analysis
Patient Global Impression of Severity–Atopic Dermatitis (PGI-S-AD)	• PGI-S-AD score • Change from baseline in PGI-S-AD score	MMRM	ITT	Bari 4 mg or Bari 2 mg vs PBO; at each post-baseline visit	Secondary Analysis
		ANCOVA using mLOCF	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Sensitivity Analysis

LY3009104

Measure	Variable	Analysis Method (Section 6.2.3)	Population (Section 6.2.1)	Comparison/Time Point	Analysis Type
Hospital Anxiety Depression Scale (HADS)	• HADS domain scores • Change from baseline in HADS domain	MMRM	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 2, 4, 8,16	Secondary Analysis
		ANCOVA using mLOCF	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Sensitivity Analysis
Dermatology Life Quality Index (DLQI)	• DLQI total score • Change from baseline in DLQI	MMRM	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Secondary Analysis

	• Observed and change from baseline in domain scores
	  -Symptoms and feelings
	  -Daily activities
	  -Leisure
	  -Work and school
	  -Personal relationships
	  -Treatment
		ANCOVA using mLOCF	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Sensitivity Analysis

	Proportion of patients achieving a 4-point improvement in DLQI total score in subgroup of patients with baseline DLQI ≥4	Logistic regression using NRI	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 1, 2, 4, 8, 16	Other Secondary Analysis
	Proportion of patients achieving DLQI (0,1)	Logistic regression using NRI	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 1, 2, 4, 8, 16	Other Secondary Analysis
Work Productivity and Activity Impairment: Atopic Dermatitis (WPAI-AD)	Observed and Change from baseline in employment status	Descriptive statistics (observed)	ITT	No comparisons; Week 16	Secondary Analysis
	Observed and Change from baseline in: • absenteeism • presenteeism • overall work impairment • impairment in activities	MMRM	ITT	Bari 4 mg or Bari 2 mg vs PBO; at each post-baseline visit	Secondary Analysis
		ANCOVA using mLOCF	ITT	Bari 4 mg or Bari 2 mg vs PBO; at each post-baseline visit	Sensitivity Analysis

Measure	Variable	Analysis Method (Section 6.2.3)	Population (Section 6.2.1)	Comparison/Time Point	Analysis Type
European Quality of Life–5 Dimensions–5 Levels (EQ-5D-5L)	Observed values in • EQ-5D mobility • EQ-5D self-care • EQ-5D usual activities • EQ-5D pain/ discomfort • EQ-5D anxiety/ depression	Logistic Regression using NRI	ITT	Bari 4 mg or Bari 2 mg vs PBO: at each post-baseline visit	Exploratory Analysis
	Observed and Change from baseline in • EQ-5D VAS • EQ-5D-5L UK Population-based index score • EQ-5D-5L US Population-based index score	MMRM	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Secondary Analysis
		ANCOVA using mLOCF	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Sensitivity Analysis
PROMIS	Observed and Change from baseline in: • PIQ – Itch Interference score • PIQ – Activity and Clothing score • PIQ – Mood and Sleep score • PIQ – Scratching Behavior score • PROMIS – Sleep-Related Impairment score	MMRM	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 1, 2, 4, 8, 12, 16	Exploratory Analysis
		ANCOVA using mLOCF	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 1, 2, 4, 8, 12, 16	Sensitivity Analysis
Neuro-QoL – Cognitive Function score	Observed and Change from baseline in: Neuro-QoL – Cognitive Function score	ANCOVA using mLOCF	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Exploratory Analysis

Measure	Variable	Analysis Method (Section 6.2.3)	Population (Section 6.2.1)	Comparison/Time Point	Analysis Type
PBI	• PBI global score • Reducing social impairments subscale score • Reducing psychological impairments subscale score • Reducing impairments due to therapy subscale score • Reducing physical impairments subscale score • Having confidence in healing subscale score	ANOVA[a]	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Exploratory Analysis
PBI	Proportion of patients achieving PBI global score ≥1 at Week 16	Logistic regression using NRI	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Exploratory Analysis

Abbreviations: ANCOVA = analysis of covariance; ANOVA= analysis of variance; Bari = baricitinib; ITT = intent-to-treat; mLOCF = modified last observation carried forward; MMRM = mixed model repeated measures; NRI = nonresponder imputation; PBO = placebo; pMI=placebo multiple imputation; PPS = per protocol set.

Notes: for all other post-baseline visits not mentioned in the table, but collected for the measures as specified in the protocol, the analyses are made as exploratory analyses.

Notes: all the key secondary are performed for Japan population. Other secondary and exploratory analyses may be performed for Japan population.

[a] ANOVA model includes region, baseline disease severity (IGA) and treatment as factors in the model.

6.14. Bioanalytical and Pharmacokinetic/Pharmacodynamic Methods

Pharmacokinetic (PK), Pharmacodynamic (PD) and Biomarker analyses to address secondary and exploratory objectives of this study will be described by Lilly in separate PK/PD and Biomarker analysis plans.

6.15. Safety Analyses

The general methods used to summarize safety data, including the definition of baseline and postbaseline are described in Section 6.2.

Safety analyses will include data from first dose of the study treatment to after rescue, unless otherwise stated, and patients will be analyzed according to the investigational product to which they were randomized at Visit 2. A sensitivity approach to the safety analyses will use data censored at last dose of the study drug for patients rescued to systemic therapy. These analyses will be conducted for treatment-emergent adverse events (TEAEs), serious adverse events (SAEs), AEs leading to permanent study drug discontinuation and special topics excluding deaths and malignancies. Additional analyses may be conducted using data after rescue to systemic therapy for some safety topics such as systemic TEAEs, and SAEs. Safety analyses will use the safety population defined in Section 6.2.1.

Safety topics that will be addressed include the following: AEs including TEAEs and SAEs, clinical laboratory evaluations, vital signs and physical characteristics, Columbia Suicide Severity Rating Scale (C-SSRS), the Self-Harm Supplement Form, safety in special groups and circumstances, including adverse events of special interest (AESI) (see Section 6.15.5), and investigational product interruptions.

Unless otherwise specified, by-visit summaries will include planned on-treatment visits. For tables that summarize events (such as AEs, categorical lab abnormalities, shift to maximum value), post-last dose follow-up data will be included. Follow-up data is defined as all data occurring up to 30 days (planned maximum follow-up time) after last dose of treatment including rescue, regardless of study period.

For selected safety assessments other than events, descriptive statistics may be presented for the last measure observed during post-treatment follow-up (up to 30 days after the last dose of treatment including rescue, regardless of study period).

6.15.1. Extent of Exposure

Duration of exposure (in days) will be calculated as follows:

- Duration of exposure to investigational product (including exposure after the initiation of rescue therapy): *date of last dose of study drug including rescue – date of first dose of study drug + 1*.

Last dose of study drug including rescue is calculated as last date on study drug. See the compound level safety standards for more details.

Total patient-years (PY) of exposure to study drug will be reported for each treatment group for overall duration of exposure. Descriptive statistics will be provided for patient-days of exposure and the frequency of patients falling into different exposure ranges in addition to cumulative exposures will be summarized.

Exposure ranges will be summarized as follows:

- ≥28 days, ≥56 days, ≥84 days, and ≥112 days
- >0 to <28 days, ≥28 days to <56 days, ≥56 days to <84 days, ≥84 days to <112 days, and ≥112 days

Overall exposure for a treatment group will be summarized in total PY which is calculated according to the following formula:

- *Exposure in PY (PYE) = sum of duration of exposure in days* (for all patients in treatment group) */ 365.25*

6.15.2. Adverse Events

Adverse events are recorded in the eCRFs. Each AE will be coded to system organ class (SOC) and preferred term (PT) using the Medical Dictionary for Regulatory Activities (MedDRA) version that is current at the time of database lock. Severity of AEs is recorded as mild, moderate, or severe.

A TEAE is defined as an event that either first occurred or worsened in severity after the first dose of study treatment and on or prior to the last visit date during the analysis period. The analysis period is defined as the treatment period plus up to 30 days off-drug follow-up time.

Adverse events are classified based upon the MedDRA PT. The MedDRA Lowest Level Term (LLT) will be used in defining which events are treatment-emergent. The maximum severity for each LLT during the baseline period up to first dose of the study medication will be used as baseline. If an event with missing severity is preexisting during the baseline period, and persists during the treatment period, then the baseline severity will be considered mild for determining treatment-emergence (that is, the event is treatment-emergent if the severity is coded moderate or severe postbaseline and not treatment-emergent if the severity is coded mild postbaseline). If an event occurring postbaseline has a missing severity rating, then the event is considered treatment-emergent unless the baseline rating is severe, in which case the event is not treatment-emergent. The day and time for events where onset is on the day of the first dose of study treatment will both be used to distinguish between pretreatment and posttreatment to derive treatment-emergence. Should there be insufficient data for AE start date to make this comparison (for example, the AE start year is the same as the treatment start year, but the AE start month and day are missing), the AE will be considered treatment-emergent.

In general, summaries will include the number of patients in the safety population (N), frequency of patients reporting the event (n), and relative frequency (that is, percentage; $n/N*100$). For any events that are gender-specific based on the displayed PT, the denominator used to compute the percentage will only include patients from the appropriate gender.

In an overview table, the number and percentage of patients in the safety population who experienced death, an SAE, any TEAE, discontinuation from the study due to an AE, permanent discontinuation from study drug due to an AE, or a severe TEAE will be summarized by treatment group.

The number and percentage of patients with TEAEs will be summarized by treatment group in 2 formats by MedDRA PT:

- nested within SOC with decreasing frequency in SOC, and events ordered within each SOC by decreasing frequency in the baricitinib 4-mg group;
- with events ordered by decreasing frequency in the baricitinib 4-mg group.

6.15.2.1. Common Adverse Events

Common TEAEs are defined as TEAEs that occurred in ≥2% (before rounding) of patients in any treatment group including placebo. The number and percentage of patients with common TEAEs will be summarized by treatment using MedDRA PT ordered by decreasing frequency in the baricitinib 4-mg group.

The number and percentage of patients with TEAEs will be summarized by maximum severity by treatment using MedDRA PT ordered by decreasing frequency in the baricitinib 4-mg group for the common TEAEs. For each patient and TEAE, the maximum severity for the MedDRA level being displayed is the maximum postbaseline severity observed from all associated LLTs mapping to that MedDRA PT.

6.15.2.2. Serious Adverse Event Analyses

Consistent with the International Conference on Harmonisation (ICH) E2A guideline (1994) and 21 Code of Federal Regulations (CFR) 312.32 (a) (2010), a SAE is any AE that results in any one of the following outcomes:

- death
- initial or prolonged inpatient hospitalization
- a life-threating experience (that is, immediate risk of dying)
- persistent or significant disability/incapacity
- congenital anomaly/birth defect

The number and percentage of patients who experienced any SAE will be summarized by treatment using MedDRA PT nested within SOC. Events will be ordered by decreasing frequency in the baricitinib 4-mg group within decreasing frequency in SOC. The SAEs will also be summarized by treatment using MedDRA PT without SOC.

An individual listing of all SAEs will be provided. A listing of deaths, if any, regardless of when they occurred during the study, will also be provided.

6.15.2.3. Other Significant Adverse Events

Other significant AEs to be summarized will provide the number and percentage of patients who:

- permanently discontinued study drug because of an AE or death

- temporarily interrupted study drug because of AE

by treatment using MedDRA PT nested within SOC. Events will be ordered by decreasing frequency in the baricitinib 4-mg group within decreasing frequency in SOC.

A summary of temporary interruptions of study drug will also be provided, showing the number of patients who experienced at least one temporary interruption and the number of temporary interruptions per patient with an interruption. Further, the duration of each temporary interruption (in days), the cumulative duration of dose interruption (in days) using basic descriptive statistics and the reason for dose interruption will be provided.

A listing of all AEs leading to permanent discontinuation from the study drug or from the study will be provided. A listing of all temporary study drug interruptions, including interruptions for reasons other than AEs, will be provided.

6.15.2.4. Criteria for Notable Patients

Patient narratives will be provided for all patients who experience certain "notable" events prior to data cutoff for the submission. See compound level safety standards for list of criteria.

6.15.3. Clinical Laboratory Evaluation

For the categorical laboratory analyses (shift and treatment emergent), the analysis period is defined as the treatment period plus up to 30 days off-drug follow-up time. The analysis period for the continuous laboratory analyses (e.g., change from baseline by time point) is defined as the treatment period excluding off-drug follow-up time.

All laboratory tests will be presented using the International Système (SI) and US conventional (CN) units. The performing central laboratory reference ranges will be used to define the low and high limits. Results pertaining to the 4 key hepatic laboratory assessments (alanine aminotransferase [ALT], aspartate aminotransferase [AST], total bilirubin, and alkaline phosphatase [ALP]) will be included as a separate analysis to address the risk of liver injury as a special safety topic (see Section 6.15.5.1).

There is one special circumstance for laboratory values to be derived based on regularly scheduled, protocol-specified analytes. The low-density lipoprotein/high-density lipoprotein (LDL/HDL) ratio will be derived as the ratio of LDL cholesterol to HDL cholesterol. There are no central lab reference ranges for the LDL/HDL ratio.

The following will be conducted for the laboratory analytes collected quantitatively:

- Box plots: Values at each visit (starting from randomization) and change from last baseline to each visit and to last postbaseline measure will be displayed in box plots for patients who have both a baseline and at least 1 postbaseline visit. The last non-missing observation in the treatment period will be used as the last observation. Individual measurements outside of reference limits will also be displayed using distinct symbols overlaying the box plot. Original-scale data will be used for the display but for some analytes (for example, immunoglobulins) a logarithmic scale may be used to aid in viewing the measures of central tendency and dispersion. Unplanned measurements will

be excluded. Descriptive summary statistics will be included below the box plot along with p-values resulting from between treatment comparison in change from last baseline to last observation. An ANCOVA model with explanatory term for treatment and the baseline value as a covariate will be used. These box plots will be used to evaluate trends over time and to assess a potential impact of outliers on central tendency summaries.

- <u>Treatment-emergent high/low analyses:</u> The number and percentage of patients with treatment-emergent high and low laboratory results at any time will be summarized by treatment group. Planned and unplanned measurements will be included. A treatment-emergent **high** result is defined as a change from a value less than or equal to the high limit at all baseline visits to a value greater than the high limit at any time during the treatment period. A treatment-emergent **low** result is defined as a change from a value greater than or equal to the low limit at all baseline visits to a value less than the low limit at any time during the treatment period. The Fisher's exact test will be used for the treatment comparisons. Number at risk (NAR) for the treatment-emergent high result is defined as the number of patients with a value less than or equal to the high limit at all baseline visits. NAR for the treatment-emergent low result is defined as the number of patients with a value greater than or equal to the low limit at all baseline visits for the treatment-emergent low result.

A listing of abnormal findings will be provided for laboratory analyte measurements, including qualitative measures. The listing will include but not limited to patient ID, treatment group, laboratory collection date, analyte name, and analyte finding. If needed by the safety physician/scientist, for analytes measured qualitatively, the number and percentage of patients with treatment-emergent abnormal laboratory results at any time will be summarized by treatment. Planned and unplanned measurements will be included. A treatment-emergent abnormal result is defined as a change from normal at all baseline visits to abnormal at any time postbaseline.

Note that additional analyses of certain laboratory analytes will be discussed within sub-sections of Section 6.15.5 pertaining to Special Safety topics (Section 6.15.5.1 for hepatic analytes, Section 6.15.5.2 for analytes related to hematological changes, Section 6.15.5.3 for analytes related to lipids, Section 6.15.5.4 for analytes related to renal function, and Section 6.15.5.5 for CPK).

6.15.4. Vital Signs and Other Physical Findings

For the treatment-emergent categorical analyses (shift and treatment emergent), the analysis period is defined as the treatment period plus up to 30 days off-drug follow-up time. The analysis period for the continuous analyses (e.g., change from baseline by time point) is defined as the treatment period excluding off-drug follow-up time.

Vital signs and physical characteristics include systolic blood pressure (SBP), diastolic blood pressure (DBP), pulse, weight, and BMI. Original-scale data will be analyzed. When these parameters are analyzed as continuous numerical variables, unplanned measurements will be

excluded. When these parameters are analyzed as categorical outcomes and/or treatment-emergent abnormalities, planned and unplanned measurements will be included.

The planned analyses described for the laboratory analytes in Section 6.15.3 will be used to analyze the vital signs and physical characteristics.

Table JAIY.6.8 defines the low and high baseline values as well as the criteria used to define treatment-emergence based on post-baseline values. The blood pressure and pulse rate criteria are consistent with the document *Selected Reference Limits for Pulse/Heart Rate, Arterial Blood Pressure (Including Orthostasis), and Electrocardiogram Numerical Parameters for Use in Analyses of Phase 2-4 Clinical Trials Version 1.3* approved on April 29, 2015 as recommended by the Lilly Cardiovascular Safety Advisory Committee (CVSAC).

Table JAIY.6.8.   **Categorical Criteria for Abnormal Treatment-Emergent Blood Pressure and Pulse Measurement, and Categorical Criteria for Weight Changes for Adults**

Parameter (Units of Measure)	Low	High
Systolic Blood Pressure (mm Hg)	≤90 (low limit) and decrease from lowest value during baseline ≥20 if >90 at each baseline visit	≥140 (high limit) and increase from highest value during baseline ≥20 if <140 at each baseline visit
Diastolic Blood Pressure (mm Hg)	≤50 (low limit) and decrease from lowest value during baseline ≥10 if >50 at each baseline visit	≥90 (high limit) and increase from highest value during baseline ≥10 if <90 at each baseline visit
Pulse (beats per minute)	<50 (low limit) and decrease from lowest value during baseline ≥15 if ≥50 at each baseline visit	>100 (high limit) and increase from highest value during baseline ≥15 if ≤100 at each baseline visit
Weight (kilograms)	(Loss) decrease ≥7% from lowest value during baseline	(Gain) increase ≥7% from highest value during baseline

6.15.5. Special Safety Topics, including Adverse Events of Special Interest

In addition to general safety parameters, safety information on specific topics of special interest will also be presented. Additional special safety topics may be added as warranted. The topics outlined in this section include the protocol-specified AESI.

In general, for topics regarding safety in special groups and circumstances, patient profiles and/or patient listings, where applicable, will be provided when needed to allow medical review of the time course of cases/events, related parameters, patient demographics, study drug treatment and meaningful concomitant medication use. In addition to the safety topics for which provision or review of patient data is specified, these will be provided when summary data are insufficient to permit adequate understanding of the safety topic.

6.15.5.1. Abnormal Hepatic Tests

Analyses for abnormal hepatic tests will involve 4 laboratory analytes: ALT, AST, total bilirubin, and ALP. In addition to the analyses described in Section 6.15.3, this section describes specific analyses for this topic.

First, the number and percentage of patients with the following abnormal elevations in hepatic laboratory tests at any time will be summarized between treatment groups:

- The percentages of patients with an ALT measurement $\geq 3\times$, $5\times$, and $10\times$ the central laboratory upper limit of normal (ULN) during the treatment period will be summarized for all patients with a postbaseline value and for subsets based on various levels of baseline.
  - The analysis of $3\times$ ULN will contain 4 subsets: patients whose non-missing maximum baseline value is $\leq 1\times$ ULN, patients whose maximum baseline is $>1\times$ ULN but $<3\times$ ULN, patients whose maximum baseline value is $\geq 3\times$ ULN, and patients whose baseline values are missing.
  - The analysis of $5\times$ ULN will contain 5 subsets: patients whose non-missing maximum baseline value is $\leq 1\times$ ULN, patients whose maximum baseline is $>1\times$ ULN but $<3\times$ ULN, patients whose maximum baseline is $\geq 3\times$ ULN but $<5\times$ ULN, patients whose maximum baseline value is $\geq 5\times$ ULN, and patients whose baseline values are missing.
  - The analysis of $10\times$ ULN will contain 6 subsets: patients whose non-missing maximum baseline value is $\leq 1\times$ ULN, patients whose maximum baseline is $>1\times$ ULN but $<3\times$ ULN, patients whose maximum baseline is $\geq 3\times$ ULN but $<5\times$ ULN, patients whose maximum baseline is $\geq 5\times$ ULN but $<10\times$ ULN, patients whose maximum baseline value is $\geq 10\times$ ULN, and patients whose baseline values are missing.
- The percentages of patients with an AST measurement $\geq 3\times$, $5\times$, and $10\times$ the central laboratory ULN during the treatment period will be summarized for all patients with a postbaseline value and for subsets based on various levels of baseline. Analyses will be constructed as described above for ALT.
- The percentages of patients with a total bilirubin measurement $\geq 2\times$ the central laboratory ULN during the treatment period will be summarized for all patients with a postbaseline value and subset into 4 subsets: patients whose non-missing maximum baseline value is $\leq 1\times$ ULN, patients whose maximum baseline is $>1\times$ ULN but $<2\times$ ULN, patients whose maximum baseline value is $\geq 2\times$ ULN, and patients whose baseline values are missing.
- The percentages of patients with an ALP measurement $\geq 1.5\times$ the central laboratory ULN during the treatment period will be summarized for all patients with a postbaseline value and subset into 4 subsets: patients whose non-missing maximum baseline value is $\leq 1\times$ ULN, patients whose maximum baseline is $>1\times$ ULN but $<1.5\times$ ULN, patients whose maximum baseline value is $\geq 1.5\times$ ULN, and patients whose baseline values are missing.

Information collected from additional hepatic safety data collection forms will be provided in patient profiles.

Second, to further evaluate potential hepatotoxicity, an Evaluation of Drug-Induced Serious Hepatotoxicity (eDISH) plot using maximum postbaseline ALT divided by ULN vs. maximum postbaseline total bilirubin divided by ULN will be created that includes all patients from the safety populations for the studies included in the submission (any phase, any medication). Each subject with at least 1 postbaseline ALT and total bilirubin contributes 1 point to the plot. The measurements do not need to be taken at the same blood draw. Symbols may be used to indicate randomized treatment.

When criteria are met for hepatic evaluation and completion of the hepatic safety CRF, investigators are required to answer a list of questions (see Compound level safety standards). A listing of the collected information will be generated together with a graphical patient profile. This includes demographics, disposition, and a display of study drug exposure, AEs, medications, and the liver-related measurements over time will be provided for these patients and any additional patients meeting ALT or AST measurement greater than or equal to 5× ULN (on a single measurement) or ALP measurement greater than or equal to 2× ULN (on a single measurement).

6.15.5.2. Hematologic Changes

Hematologic changes will be defined based on clinical laboratory assessments. Common Terminology Criteria for Adverse Events (CTCAEs) will be applied for selected laboratory tests and are described in the compound level safety standards. These CTCAE grading schemes are consistent with both Version 3.0 and Version 4.03 of the CTCAE guidelines (CTCAE 2003, 2010).

Treatment-emergent laboratory abnormalities occurring at any time during the treatment period and shift tables of baseline to maximum grade during the treatment period will be tabulated. Planned and unplanned measurements will be included. Treatment-emergence will be characterized using the following 5 criteria (as appropriate to the grading scheme):

- any increase in postbaseline CTCAE grade from worst baseline grade
- increase to Grade 1 or above at worst postbaseline
- increase to Grade 2 or above at worst postbaseline
- increase to Grade 3 or above at worst postbaseline
- increase to Grade 4 at worst postbaseline.

Shift tables will show the number and percentage of patients based on baseline to maximum during the treatment period, with baseline depicted by the most extreme grade during the baseline period. With each shift table, a shift table summary displaying the number and percentage of patients with maximum postbaseline results will be presented by treatment group for each treatment period within the following categories:

- decreased: postbaseline category < baseline category
- increased: postbaseline category > baseline category

- same: postbaseline category = baseline category

A laboratory-based treatment-emergent outcome related to increased platelet count will be summarized in similar fashion. Treatment-emergent thrombocytosis as a laboratory-based abnormality will be defined as an increase in platelet count from a maximum baseline value ≤600 billion/L to any postbaseline value >600 billion/L (Lengfelder et al. 1998). Planned and unplanned measurements will be included.

A listing of patients with treatment-emergent thrombocytosis may be provided for safety review.

6.15.5.3. Lipids Effects

Lipid effects will be assessed through analysis of elevated total cholesterol, elevated LDL cholesterol, decreased HDL cholesterol, and elevated triglycerides as described in Section 6.15.3 and with TEAEs potentially related to hyperlipidemia.

Categorical analyses will be performed using National Cholesterol Education Program (NCEP) Adult Treatment Panel (ATP) III guidelines (2002) as shown in the compound level safety standards. The grade-like categories shown in this table are ordered from traditionally most desirable to least desirable for the purposes of these analyses.

Shift tables will show the number and percentage of patients based on baseline to the least desirable category during the treatment period, with baseline depicted by the least desirable category during the baseline period. With each shift table, a shift table summary displaying the number and percentage of patients with the least desirable postbaseline results will be presented by treatment group for each treatment period within the following categories:

- decreased: postbaseline category more desirable than baseline category,
- increased: postbaseline category less desirable than baseline category,
- same: postbaseline category = baseline category.

Treatment-emergent laboratory abnormalities related to elevated total cholesterol, elevated triglycerides, elevated LDL cholesterol, and decreased and increased HDL cholesterol occurring at any time during the treatment period will be tabulated using the NCEP categories shown in the compound level safety standards.

Treatment-emergent elevated total cholesterol will be characterized as follows:

- increase to categories 'Borderline high' or 'High'
- increase to category 'High.'

Treatment-emergent elevated triglycerides will be characterized as

- increase to categories 'Borderline high,' 'High,' or 'Very high'
- increase to categories 'High' or 'Very high'
- increase to category 'Very high.'

Treatment-emergent elevated LDL cholesterol will be characterized as

- increase to categories 'Borderline high,' 'High,' or 'Very high'

LY3009104

- increase to categories 'High' or 'Very high'
- increase to 'Very high'

Treatment-emergent abnormal HDL cholesterol will be characterized as

- decreased HDL
    - decrease to categories 'Normal' or 'Low'
    - decrease to category 'Low'
- increased HDL
    - increase to categories 'Normal' or 'High'
    - increase to category 'High'

The percentages of patients with treatment-emergent potential hyperlipidemia will be summarize by treatment, ordered by decreasing frequency in the baricitinib 4-mg group using a predefined MedDRA list of PTs that is a subset of the narrow scope PTs in the MedDRA SMQ 'Dyslipidemia' (code 200000026) [see Compound level safety standards].

6.15.5.4. Renal Function Effects

Effects on renal function will be assessed through analysis of elevated creatinine.

CTCAEs will be applied for laboratory tests related to renal effects as shown in the compound level safety standards. This CTCAE grading scheme is consistent with both Version 3.0 and Version 4.03 of the CTCAE guidelines. Shift tables will show the number and percentage of patients based on baseline to maximum during the treatment period, with baseline depicted by highest grade during the baseline period. Treatment-emergent laboratory abnormalities related to elevated creatinine occurring at any time during the analysis period will be tabulated. Refer to the Compound level safety standards for details.

6.15.5.5. Elevations in Creatine Phosphokinase (CPK)

Elevations in creatine phosphokinase (CPK) will be addressed using CTCAE criteria as described in the compound level safety standards. This CTCAE grading scheme is consistent with both Version 3.0 and Version 4.03 of the CTCAE guidelines. Analyses will be the same as the CTCAE analyses specified for laboratory tests related to renal function events in Section 6.15.5.2.

A listing of elevated CPK (CTCAE grade of 3 or above) will be provided for medical safety review.

Treatment-emergent adverse events potentially related to muscle symptoms may be analyzed, based on reported AEs. The Muscle Symptoms special search category is a pre-defined MedDRA search criteria list that contains the narrow scope terms from the Rhabdomyolysis / myopathy SMQ (code 20000002) plus selected terms from the Musculoskeletal SOC. These terms are shown in the compound level safety standards.

6.15.5.6. Infections

Infections will be defined using all the PTs from the Infections and Infestations SOC as defined in MedDRA. Serious infection will be defined as all the infections that meet the SAE criteria.

LY3009104

The number and percentage of patients with TEAEs of infections, serious infections, and infections resulting in permanent study drug discontinuation will be summarized by treatment group using MedDRA PTs. The proportion of patients developing skin infections requiring antibiotic treatment by Week 16 will also be summarized on the overview of infections table.

The number and percentage of patients with TEAEs of infections by maximum severity will be summarized by treatment group using MedDRA PTs.

Treatment-emergent infections will be reviewed in context of other clinical and laboratory parameters via a listing (details see Compound level safety standards).

The TEAE infections will be further analyzed in terms of potential opportunistic infection, herpes zoster and herpes simplex. Summary of HBV DNA monitoring results and association between infection and neutropenia/lymphopenia will also be provided in the context of infections.

Opportunistic infection

To identify potential opportunistic infections (POIs), the following approach will be used:

- identifying the POIs using a list of MedDRA PTs (refer to the compound level safety standards).

Potential opportunistic infections identified through these search approaches may be combined in one list for medical assessment and final classification of whether the case met the modified Winthrop definitions for opportunistic infections (OIs).

A final listing for OIs will be provided for the CSR and to assist the composition of patient narratives.

Herpes zoster

Cases of herpes zoster will be further classified as follows:

- localized or non-multidermatomal-involvement of the primary and/or adjacent dermatomes only
    - complicated – documented ocular (cornea or deeper structure; for example, iritis, keratitis, retinitis, etc.) or motor nerve involvement (e.g., palsy; post herpetic neuralgia [PHN] does not meet criteria for motor nerve involvement).
    - uncomplicated-localized or non-multidermatomal cases that are not complicated
- multidermatomal-involvement beyond primary and adjacent dermatomes (that is, >3 contiguous dermatomes) or involvement of two or more non-contiguous dermatomes
    - complicated-documented ocular (cornea or deeper structure; for example iritis, keratitis, retinitis, etc.) or motor nerve involvement
    - uncomplicated-multidermatomal cases
- disseminated-systemic infection, visceral or widespread cutaneous (e.g., ≥5 dermatomes or 3 to 4 dermatomes including at least 1 non-contiguous).

- Recurrent - >1 infection occurring in an individual patient during the course of participation in the baricitinib clinical program.

All herpes zoster will undergo medical review to determine the classification as described above.

A summary of herpes zoster table will be provided. The summary table will also include event maximum severity, seriousness, whether resulting in temporary study drug interruption, whether resulting in study drug discontinuation, whether treated with antiviral medication, and event outcome. The incidence rate adjusted for observation time will also be provided (as defined in Section 6.15.2). Of note, in the context of herpes zoster, antiviral medication treatment is defined as that the medication was initiated at the event start date, or within 30 days before or after the event start date. The antiviral medication for herpes zoster includes but not limited to Aciclovir, Brivudine, Cidofovir, Famciclovir, Foscarnet, Ganciclovir, Penciclovir, Valaciclovir, Valganciclovir, Vidarabine (best presented by J05AB, J05AC, J05AE, and J05AH Anatomical Therapeutic Classification codes). Medical representatives will review the concomitant medication list prior to database lock and make adjustment of the above list if necessary.

If a patient has more than 1 event of herpes zoster, the event with the maximum severity will be used in these summary tables. If more than 1 event of herpes zoster occurs with the same severity, the event with the longest duration will be used in the summary table.

Herpes simplex

A summary analysis of herpes simplex will be provided. Herpes simplex will be defined based on MedDRA PT as listed in the compound level safety standards (both narrow and broad terms in the herpes simplex section). The list needs to be reviewed by medical prior to data locks (final and interim). The summary table will include event maximum severity, seriousness, whether resulting in temporary study drug interruption, whether resulting in study drug discontinuation, and whether treated with antiviral medication.

If a patient has more than 1 event of herpes simplex, the event with the maximum severity will be used in these summary tables. If more than 1 event of herpes simplex occurs with the same severity, the event with the longest duration will be used in the summary table.

Skin Infections

A summary analysis of skin infections may be provided. Skin infections may be defined based on MedDRA preferred term (see the Compound level safety standards).

HBV DNA

A listing of patients with detectable hepatitis B virus (HBV) deoxyribonucleic acid (DNA) post baseline will be provided.

HBV DNA status post baseline (not detectable, detectable but not quantifiable [i.e., < lower limit of detection (LLOD)], quantifiable [i.e., ≥LLOD]) will be summarized by treatment group stratified by baseline HBV serology status, specifically:

- HBsAb+ / HBcAb+

- HBsAb- / HBcAb+

Association between infection and neutropenia/lymphopenia

Depending on the number of cases with CTCAE Grade 2 or greater, a summary table will be provided for treatment-emergent infections that were preceded or accompanied by neutropenia. For this analysis, neutropenia is defined as (1) CTCAE Grade 2 or greater, (2) Grade 3 or greater. Infection events with onset date ≤14 days before or after the Grade 2 or greater neutrophil count collection date will be considered as infections preceded or accompanied by neutropenia.

Similar analyses as above will be conducted to evaluate the association between infection and lymphopenia.

6.15.5.7. Major Adverse Cardiovascular Events (MACE) and Other Cardiovascular Events

Potential major adverse cardiovascular events (MACE) and other cardiovascular events requiring adjudication will be analyzed.

Categories and subcategories analyzed will include, but are not limited to, the following:

- MACE
    - o Cardiovascular death,
    - o Myocardial infarction (MI),
    - o Stroke,
- Other cardiovascular events
    - o Transient ischemic attack,
    - o Hospitalization for unstable angina,
    - o Hospitalization for heart failure,
    - o Serious arrhythmia,
    - o Resuscitated sudden death,
    - o Cardiogenic shock,
    - o Coronary interventions (such as coronary artery bypass surgery or percutaneous coronary intervention),
- Non-cardiovascular death,
- All-cause death.

In general, events requiring adjudication are documented by investigative sites using an endpoint reporting CRF. This CRF is then sent to the adjudication center for external adjudication which uses an adjudication reporting CRF to document the final assessment of the event as a MACE, as some other cardiovascular event, or as no event (according to the Clinical Endpoint Committee Charter). In some cases, however, the investigator may not have deemed that an event had met the endpoint criteria but the event was still sent for adjudication as a potential MACE, other cardiovascular event, or no event. These events are included in the adjudication process to ensure adequate sensitivity. In these instances, the adjudication reporting CRF will not have a matching endpoint reporting CRF from the investigator. Events generated from these

circumstances will be considered as events sent for adjudication in the absence of an investigator's endpoint reporting form.

The number and percentage of patients with MACE, other cardiovascular events, non-cardiovascular death, and all-cause death, as positively adjudicated, will be summarized by treatment group based on the categories and subcategories above.

A listing of the events sent for adjudication will be provided to include data concerning the MedDRA PT related to the event, the seriousness of the event, and the event outcome, along with the adjudicated result.

6.15.5.8. Venous and Pulmonary Artery Thromboembolic (VTE) Events

Events identified as representative of venous thromboembolic events (VTE) disease will be classified as Deep Vein Thrombosis (DVT), pulmonary embolism (PE), or other peripheral venous thrombosis and will be analyzed. The following definitions apply:

- DVT: Clinical diagnosis of a thrombosis in a deep vein above the knee that must be confirmed by objective evidence of either: a filling defect of deep veins of the leg on venography or a non-compressible venous segment on ultrasound or confirmation by other imaging modality (e.g., Computed tomography [CT], Magnetic Resonance Imaging [MRI]).
- PE: Clinical diagnosis of pulmonary embolus that must be confirmed by objective evidence of either: a filling defect of pulmonary arteries by either pulmonary angiography or CT angiography or by a high probability Ventilation Perfusion (VQ) scan.
- Other Peripheral Venous Thrombosis: Clinical diagnosis of a venous thrombosis not specified by either DVT or PE above. Other peripheral venous thrombosis disease must be confirmed by objective evidence by imaging including venography, ultrasound, CT scan, or MRI. Examples of these would include non-superficial below knee thrombosis, portal vein, subclavian vein, or mesenteric vein. Superficial thrombophlebitis alone is not considered a VTE event.

In general, events requiring adjudication are documented by investigative sites using an endpoint reporting CRF. Refer to Section 6.15.5.7 for more details as the process is the same as that of MACE.

The number and percentage of patients with a VTE, DVT/PE, DVT, PE, and other peripheral venous thrombosis, as positively adjudicated, will be summarized by treatment group. Note that the below knee thrombosis captured in the "other peripheral venous thrombosis" category will be summarized within DVT.

A listing of the VTE events sent for external adjudication will be provided to include data concerning the MedDRA PT related to the event, the seriousness of the event, and the event outcome, along with the adjudicated result.

6.15.5.9. Arterial Thromboembolic (ATE) Events

Refer to the Compound level safety standards.

6.15.5.10.   Malignancies

Malignancies will be identified using terms from the malignant tumors SMQ (SMQ 20000194). Malignancies excluding non-melanoma skin cancers (NMSC) and NMSC will be reported separately.

All the cases identified by malignant tumors SMQ will be assessed through medical review to determine confirmed NMSC cases.

First, a listing including all the malignancy cases will be prepared before database lock along with the *planned* NMSC flag according to the current MedDRA version PTs (the list will be updated depending on the MedDRA version used for analysis):

- Squamous cell carcinoma of skin (10041834)
- Bowen's disease (10006059)
- Basal cell carcinoma (10004146)
- Basosquamous carcinoma (10004178)
- Basosquamous carcinoma of skin (10004179)
- Squamous cell carcinoma (10041823)
- Skin squamous cell carcinoma metastatic (10077314)
- Skin cancer (10040808)
- Carcinoma in situ of skin (10007390)
- Keratoacanthoma (10023347)
- Vulvar squamous cell hyperplasia (10079905)
- Skin squamous cell carcinoma recurrent (10081136)

This internal review is to occur prior to database lock. The case review and subsequent summary analyses will include all the cases reported in the study database or by LSS report, disregarding the length of gap between the last treatment dose date and the event date. The NMSC flag will be confirmed during the internal review process.

The number and percentage of patients with TEAEs associated malignancies excluding NMSC and NMSC will be summarized by treatment group.

6.15.5.11.   Allergic Reactions/Hypersensitivities

A search will be performed using the MedDRA version 21.1 SMQs to search for relevant events, using the following queries:

- Anaphylactic reaction SMQ (20000021)
- Hypersensitivity SMQ (20000214)
- Angioedema SMQ (20000024)

Assessment of the Anaphylactic reaction SMQ includes an algorithmic query. The algorithmic approach comprises one or more events associated with an individual administration of study drug, where the events include:

- A narrow term from the SMQ (Category A of the SMQ);
- Multiple terms from the SMQ, comprising terms from at least two of the following categories from the SMQ:
  - Category B - (Upper Airway/Respiratory signs and symptoms)
  - Category C - (Angioedema/Urticaria/Pruritus/Flush signs and symptoms)
  - Category D - (Cardiovascular/Hypotension signs and symptoms).

Refer to the Compound safety level standards for details.

6.15.5.12. Gastrointestinal Perforations

Treatment-emergent adverse events related to potential gastrointestinal (GI) perforations will be analyzed using reported AEs. Identification of these events will be based on review of the PTs of the MedDRA SMQ 20000107, GI perforations (note that this SMQ holds only narrow terms and has no broad terms). Potential GI perforations identified by the above SMQ search will be provided as a listing for internal review. Each case will be assessed to determine whether it is GI perforation. A summary table based on medical review may be provided and treatment comparisons will be made using Fisher's exact test.

6.15.5.13. Columbia Suicide Severity Rating Scale

Suicidal ideation, suicidal behavior, and self-injurious behavior without suicidal intent, based on the C-SSRS, will be listed by patient and visit. Only patients that show suicidal ideation/behavior or self-injurious behavior without suicidal intent during treatment will be displayed along with all their ideation and behavior, even if not positive (i.e., if a patient's answers are all 'no' for the C-SSRS, then that patient will not be displayed). A summary of the C-SSRS categories during treatment and a shift summary in the C-SSRS categories from baseline during treatment will be provided. Refer to the Compound safety level standards for details.

6.15.5.14. Self-Harm Supplement Form and Self-Harm Follow-up Form

The Self-Harm Supplement Form is a single question to enter the number of suicidal behavior events, possible suicide behaviors, or nonsuicidal self-injurious behaviors. If the number of behavioral events is greater than zero, it will lead to the completion of the Self-Harm Follow-Up Form. The Self-Harm Follow-Up Form is a series of questions that provides a more detailed description of the behavior cases. A listing of the responses give on the Self-Harm Follow-Up Form will be provided.

6.16. Subgroup Analyses

Subgroup analyses comparing each dose of baricitinib to placebo will be performed on the ITT population at Week 16 using the primary censoring rule for the following:

- Proportion of patients achieving IGA 0 or 1

- Proportion of patients achieving EASI75 Response Rate

- Proportion of patients achieving Itch NRS 4-point improvement

The following subgroups, categorized into disease-related characteristics and demographic characteristics, will be evaluated:

- Patient Demographic and Characteristics Subgroups:
  - Gender (male, female)
  - Age group (<65, ≥65 years old)
  - Age group (<65, ≥65 to <75, ≥75 to <85, ≥85 years old)
  - Baseline weight: (<60 kg, ≥60 to <100 kg, ≥100 kg)
  - Baseline BMI (<25 kg/m^2, ≥25 to <30 kg/m^2, ≥30 kg/m^2)
  - Race: (American Indian or Alaska Native, Asian, Black or African American, Native Hawaiian or Other Pacific Islander, White, Multiple)
  - Baseline renal function status: impaired (eGFR <60 mL/min/1.73 m^2) or not impaired (eGFR ≥60 mL/min/1.73 m^2)
- Geographic Region Subgroups:
  - Region: (as defined in Table JAIY.5.1)
  - Specific regions (Europe, other)
  - Specific regions (East Asia[Korea, Japan and Taiwan], other)
  - Specific country (Japan, other)
  - Prior systemic therapy use (yes, no)
- Baseline Disease-Related Characteristics Subgroup
  - Baseline disease severity (IGA score): 3, 4

Descriptive statistics will be provided for each treatment and stratum of a subgroup as outlined, regardless of sample size. As all endpoints are categorical, subgroup analyses will be performed using logistic regression using Firth's correction to accommodate (potential) sparse response rates. The model will include the categorical outcome as the dependent variable and baseline value (for EASI and itch), baseline severity, treatment, subgroup, and treatment-by-subgroup interaction as explanatory variables. Missing data will be imputed using NRI (Section 6.4.1). The treatment-by-subgroup interaction comparing treatment groups will be tested at the 0.1 significance level. The p-value from the logistic regression model will be reported for the interaction test and the subgroup test, unless the model did not converge. Response counts and percentages will be summarized by treatment for each subgroup category. The difference in percentages and 95% CI of the difference in percentages using the Newcombe-Wilson without continuity correction will be reported. The corresponding p-value from the Fisher's exact test will also be produced.

In case any level of a subgroup comprises $<10\%$ of the overall sample size, only descriptive summary statistics will be provided for treatment arms, and no treatment group comparisons will be performed within these subgroup levels.

Additional subgroup analyses on efficacy may be performed as deemed appropriate and necessary.

6.17. Protocol Deviations

Protocol deviations will be tracked by the clinical team, and their importance will be assessed by key team members during protocol deviation review meetings. Out of all important protocol deviations (IPDs) identified, a subset occurring during Period 2 with the potential to affect efficacy analyses will result in exclusion from the PP population.

Potential examples of deviations include patients who receive excluded concomitant therapy, significant non-compliance with study medication (<80% of assigned doses taken, failure to take study medication and taking incorrect study medication), patients incorrectly enrolled in the study, and patients whose data are questionable due to significant site quality or compliance issues. Refer to a separate document for the important protocol deviations.

Trial Issue Management Plan includes the categories and subcategories of important protocol deviations and whether or not these deviations will result in the exclusion of patients from per protocol set.

The number and percentage of patients having IPD(s) will be summarized within category and subcategory of deviation by treatment group for Period 2 using the ITT population. Individual patient listings of IPDs will be provided. A summary of reasons patients were excluded from the PPS will be provided by treatment group.

6.18. Interim Analyses and Data Monitoring

An interim analyses may be conducted at the time when the last patient completes Visit 8 (Week 16) or ETV.

The baricitinib AD, AA and SLE Phase 3 programs Data Monitoring Committee (DMC) is an independent expert advisory group commissioned and charged with the responsibility of evaluating cumulative safety at regular intervals. As such, the primary objective of the DMC is to monitor the safety of the subjects enrolled in the baricitinib AD, AA and SLE Phase 3 programs by reviewing the available clinical data at scheduled time points, as described in this DMC Charter, as well as on an ad hoc basis, as needed. The DMC will consist of members external to Lilly. This DMC will follow the rules defined in the DMC charter, focusing on potential and identified risks for this molecule and for this class of compounds. Data Monitoring Committee membership will include, at a minimum, specialists with expertise in dermatology, statistics, cardiology, and other appropriate specialties.

The DMC will be authorized to review unblinded results of analyses by treatment group prior to database lock, including study discontinuation data, AEs including SAEs, clinical laboratory data, vital sign data, etc. The DMC may recommend continuation of the study, as designed; temporary suspension of enrollment; or the discontinuation of a particular dose regimen or the entire study. While the DMC may request to review efficacy data to investigate the benefit/risk relationship in the context of safety observations for ongoing patients in the study, no information regarding efficacy will be communicated. Moreover, the study will not be stopped for positive efficacy results nor will it be stopped for futility. Hence, no alpha is spent. Details of the DMC, including its operating characteristics, are documented in the Baricitinib Atopic

Dermatitis DMC charter and further details are given in the Interim Analysis Plan in Section 6.18.1.

Besides DMC members, a limited number of pre-identified individuals may gain access to the limited unblinded data, as specified in the unblinding plan, prior to the interim or final database lock, for preparation of regulatory documents. Information that may unblind the study during the analyses will not be reported to study sites or blinded study team until the study has been unblinded.

6.18.1. Interim Analysis Plan

The final analyses for this study are considered as the analyses to be presented to the DMC.

6.19. Planned Exploratory Analyses

The planned exploratory analyses are described in Sections 6.12 and 6.13. Additional exploratory analyses may be conducted such as exploring inadequate or super responders and their baseline characteristics and will be documented in a supplemental SAP. Health Technology Assessment (HTA) toolkit analyses, which may be produced, will also be documented in the supplemental SAP.

6.20. Annual Report Analyses

Annual report analyses, such as the Development Update Safety Report (DSUR), will be documented in a separate document.

6.21. Clinical Trial Registry Analyses

Additional analyses will be performed for the purpose of fulfilling the Clinical Trial Registry (CTR) requirements.

Analyses provided for the CTR requirements include a summary of AEs, provided as a dataset which will be converted to an XML file. Both SAEs and 'Other' AE are summarized: by treatment group, by MedDRA PT.

- An AE is considered 'Serious' whether or not it is a TEAE.
- An AE is considered in the 'Other' category if it is both a TEAE and is not serious. For each SAE and 'Other' AE, for each term and treatment group, the following are provided:
  - the number of participants at risk of an event
  - the number of participants who experienced each event term
  - the number of events experienced.
- Consistent with www.ClinicalTrials.gov requirements, 'Other' AEs that occur in fewer than 5% of patients/subjects in every treatment group may not be included if a 5% threshold is chosen (5% is the minimum threshold).
- AE reporting is consistent with other document disclosures for example, the CSR, manuscripts, and so forth.

Similar methods will be used to satisfy the European Clinical Trials Database (EudraCT) requirements.

7. Unblinding Plan

A separate JAIY Blinding / Unblinding Plan contains details of how the blind is maintained for this study.

8.   References

Alosh M, Bretz F, Huque M. Advanced multiplicity adjustment methods in clinical trials. *Stat Med.* 2014;33(4):693-713.

Basra MK, Salek MS, Camilleri L, Sturkey R, Finlay AY. Determining the minimal clinically important difference and responsiveness of the Dermatology Life Quality Index (DLQI): further data. *Dermatology.* 2015;230(1):27-33.

Bretz F, Posch M, Glimm E, Klinglmueller F, Maurer W, Rohmeyer K. Graphical approaches for multiple comparison procedures using weighted Bonferroni, Simes, or parametric tests. *Biom J.* 2011;53(6):894-913.

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

[CTCAE] Common Terminology Criteria for Adverse Events, Cancer Therapy Evaluation Program, Version 3.0, DCTD, NCI, NIH, DHHS, March 31, 2003. Available at: https://ctep.cancer.gov/protocoldevelopment/electronic_applications/docs/ctcaev3.pdf. Accessed 25 July 2017.

[CTCAE] Common Terminology Criteria for Adverse Events, Version 4.03, DHHS NIH NCI, NIH Publication No. 09-5410, June 14, 2010. Available at: https://www.eortc.be/services/doc/ctc/CTCAE_4.03_2010-06-14_QuickReference_5x7.pdf. Accessed 25 July 2017.

EuroQol Group. EQ-5D-5L User Guide. Version 2.1. Available at: http://www.euroqol.org/fileadmin/user_upload/Documenten/PDF/Folders_Flyers/EQ-5D-5L_UserGuide_2015.pdf. Accessed March 2018.

Gillings D, Koch G. The application of the principle of intention-to-treat to the analysis of clinical trials. *Drug Information Journal.* 1991;25(3):411-424.

Hanifin JM, Thurston M, Omoto M, Cherill R, Tofte SJ, Graeber M. The eczema area and severity index (EASI): assessment of reliability in atopic dermatitis. EASI Evaluator Group. *Exp Dermatol.* 2001;10(1):11-18.

Herdman M, Gudex C, Lloyd A, Janssen MF, Kind P, Parkin D, Bonsel G, Badia X. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). *Qual Life Res.* 2011;20(10):1727-1736.

Hongbo Y, Thomas CL, Harrison MA, Salek MS, Finlay AY. Translating the science of quality of life into practice: What do dermatology life quality index scores mean? *J Invest Dermatol.* 2005;125(4):659-664.

[ICH] International Conference on Harmonisation. Harmonised Tripartite Guideline: Clinical safety data management: definitions and standards for expedited reporting. E2A. 1994;Step 4. Available at: http://www.ich.org/products/guidelines/efficacy/article/efficacy-guidelines.html. Accessed 21 July 2017.

[ICH E9 R1] Estimands and Sensitivity Analysis in Clinical Trials
http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E9/E9-R1EWG_Step2_Guideline_2017_0616.pdf

Khilji FA, Gonzalez M, Finlay AY. Clinical meaning of change in Dermatology Life Quality Index scores. *Br J Dermatol*. 2002;147(suppl 62):25-54.

Kimball AB, Naegeli AN, Edson-Heredia E, Lin CY, Gaich C, Nikai E, Yosipovitch G. Psychometric properties of the Itch Numeric Rating Scale in patients with moderate-to-severe plaque psoriasis. *Br J Dermatol.* 2016;175(1):157-162.

Kunz B, Oranje A, Labrèze, Stadler JF, Ring J, Taïeb A, Clinical Validation and Guidelines for the SCORAD Index: Consensus Report of the European Task Force Atopic Dermatitis. *Dermatology*. 1997;195(1):10-19.

Lengfelder E, Hochhaus A, Kronawitter U, Hoche D, Queisser W, Jahn-Eder M, Burkhardt R, Reiter A, Ansari H, Hehlmann R. Should a platelet limit of 600 x 10(9)/l be used as a diagnostic criterion in essential thrombocythaemia? An analysis of the natural course including early stages. *Br J Haematol*. 1998;100(1):15-23.

Naegeli AN, Flood E, Tucker J, Devlen J, Edson-Heredia E. The Worst Itch Numeric Rating Scale for patients with moderate to severe plaque psoriasis or psoriatic arthritis. *Int J Dermatol.* 2015;54(6):715-722.

[NCEP] National Cholesterol Education Program. Third Report of the NCEP Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel [ATP] III) Final Report, National Cholesterol Education Program, National Heart, Lung, and Blood Institute, National Institutes of Health, NIH Publication No. 02-5215, September 2002. *Circulation*. 2002;106(25):3143-421.

Program Safety Analysis Plan: Baricitinib (LY3009104) Version 5. Report on file, Eli Lilly and Company.

Protocol I4V-MC-JAIY(a). A Multicenter, Randomized, Double-Blind, Placebo-Controlled, Phase 3 Study to Evaluate the Efficacy and Safety of Baricitinib in Adult Patients with Moderate to Severe Atopic Dermatitis. Report on file, Eli Lilly and Company.

Reilly MC, Zbrozek AS, Dukes EM. The validity and reproducibility of a work productivity and activity impairment instrument. *Pharmacoeconomics*. 1993;4(5):353-365.

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

Snaith RP. The Hospital Anxiety and Depression Scale. *Health and Quality of Life Outcomes*. 2003;1:29.

Stalder J, Taieb A, Atherton D, Bieber P, Bonifazi E, Broberg A, Calza A, Coleman R, De Prost Y, Stalder J, Gelmetti C, Cuannetti A, Harper J, Kunz B, Lachapelle J, Langeland T, Lever R, Oranje A, Oueille-Roussel C, Revuz J, Ring J, Roujeau J, Saurat J, Song M, Tennstedt D, Van Neste D, Vieluf D, Poncet M. Severity scoring of atopic dermatitis: the SCORAD index. Consensus Report of the European Task Force on Atopic Dermatitis. *Dermatology*. 1993;186(1):23-31.

US National Archives and Records Administration. Code of Federal Regulations (CFR). Investigational New Drug Application (IND) safety reporting. 2010;Title 21:Section 312.32. Available at: https://www.ecfr.gov/cgi-bin/text-idx?SID=fd87d5f6fb00b95672dbe924a6d3b2f0&mc=true&node=se21.5.312_132&rgn=div8. Accessed 21 July 2017.

White D, Leach C, Sims R, Atkinson M, Cottrell D. Validation of the Hospital Anxiety and Depression Scale for use with adolescents. *Br J Psychiatry*. 1999;175(5):452-454.

Winthrop KL, Novosad SA, Baddley JW, Calabrese L, Chiller T, Polgreen P, Bartalesi F, Lipman M, Mariette X, Lortholary O, Weinblatt ME, Saag M, Smolen J. Opportunistic infections and biologic therapies in immune-mediated inflammatory diseases: consensus recommendations for infection reporting during clinical trials and postmarketing surveillance. *Ann Rheum Dis*. 2015;74(12):2107-2116.

Zigmond AS, Snaith RP. The Hospital Anxiety and Depression Scale. *Acta Psychiatr Scand*. 1983;67(6):361-370.