Statistical Analysis Plan v2- I4V-MC-JAIW

A Multicenter, Randomized, Double-Blind, Placebo Controlled, Phase 3 Study to Evaluate the Efficacy and Safety in Adult Patients with Moderate to Severe Atopic Dermatitis

NCT03435081

Approved Date: 13-Dec-2019

1. Statistical Analysis Plan: I4V-MC-JAIW: 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

Breeze-AD5

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-JAIW (NCT03334396) is a Phase 3, multicenter, randomized, double-blind, placebo-controlled, parallel group, outpatient, 108-week study designed to evaluate the efficacy and safety of baricitinib 1-mg and 2-mg in patients with moderate to severe atopic dermatitis.

Eli Lilly and Company
Indianapolis, Indiana USA 46285
Protocol I4V-MC-JAIW
Phase 3

Statistical Analysis Plan Version 1 electronically signed and approved by Lilly: 09 August 2018
Statistical Analysis Plan Version 2 electronically signed and approved by Lilly on date provided below

Approval Date: 13-Dec-2019 GMT

2. Table of Contents

Section **Page**

1. Statistical Analysis Plan: I4V-MC-JAIW: 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 1
2. Table of Contents 2
3. Revision History 7
4. Study Objectives 8
4.1. Primary Objective 8
4.2. Secondary Objectives 8
4.2.1. Key Secondary Objectives 8
4.2.2. Other Secondary Objectives 9
4.2.3. Other Secondary Objectives for Week 16 Responders 10
4.3. Exploratory Objectives 10
5. Study Design 12
5.1. Method of Assignment to Treatment 13
6. A Priori Statistical Methods 14
6.1. Determination of Sample Size 14
6.2. General Considerations 14
6.2.1. Reporting Periods 14
6.2.2. Analysis Populations 15
6.2.3. Definition of Baseline and Post-Baseline Measures 16
6.2.4. Analysis Methods 18
6.2.5. Derived Data 19
6.3. Adjustments for Covariates 21
6.4. Handling of Dropouts or Missing Data 21
6.4.1. Non-Responder Imputation 22
6.4.2. Mixed Model for Repeated Measures 23
6.4.3. Modified Last Observation Carried Forward 23
6.4.4. Modified Baseline Observation Carried Forward 23
6.4.5. Placebo Multiple Imputation 24
6.4.6. Tipping Point Analyses 25
6.5. Multicenter Studies 27
6.6. Multiple Comparisons/Multiplicity 27
6.7. Patient Disposition 30
6.8. Patient Characteristics 30

6.8.1. Demographics 30
6.8.2. Baseline Disease Characteristics 31
6.8.3. Historical Illness and Pre-existing Conditions 32
6.9. Treatment Compliance 32
6.9.1. Rescue Treatment 33
6.10. Previous and Concomitant Therapy 33
6.11. Efficacy Analyses 34
6.11.1. Primary Outcome and Methodology 42
6.11.2. Secondary and Exploratory Efficacy Analyses 42
6.11.3. Sensitivity Analyses 43
6.12. Health Outcomes/Quality-of-Life Analyses 43
6.13. Bioanalytical and Pharmacokinetic/Pharmacodynamic Methods 63
6.14. Safety Analyses 63
6.14.1. Extent of Exposure 64
6.14.2. Adverse Events 65
6.14.2.1. Common Adverse Events 66
6.14.2.2. Serious Adverse Events 66
6.14.2.3. Other Significant Adverse Events 67
6.14.2.4. Criteria for Notable Patients 67
6.14.3. Clinical Laboratory Evaluation 67
6.14.4. Vital Signs and Other Physical Findings 69
6.14.5. Special Safety Topics Including Adverse Events of Special Interest 69
6.14.5.1. Abnormal Hepatic Tests 70
6.14.5.2. Hematologic Changes 71
6.14.5.3. Lipids Effects 72
6.14.5.4. Renal Function Effects 73
6.14.5.5. Elevations in Creatinine Phosphokinase 73
6.14.5.6. Infections 74
6.14.5.7. Major Cardiovascular Events and other Cardiovascular Events 76
6.14.5.8. Venous and Pulmonary Artery Thromboembolic Events 77
6.14.5.9. Arterial Thromboembolic (ATE) Events 78
6.14.5.10. Malignancies 78
6.14.5.11. Allergic Reactions/Hypersensitivity 79
6.14.5.12. Gastrointestinal Perforations 79
6.14.5.13. Columbia Suicide Severity Rating Scale 79
6.14.5.13.1. Self-Harm Supplemental Form and Self-Harm Follow-up Form 80

6.15. Subgroup Analyses 80
6.16. Protocol Violations 81
6.17. Interim Analyses and Data Monitoring 81
6.17.1. Interim Analysis Plan 82
6.18. Planned Exploratory Analyses 83
6.19. Annual Report Analyses 83
6.20. Clinical Trial Registry Analyses 83
7. Unblinding Plan 84
8. References 85

Table of Contents

Table **Page**

Table JAIW.6.1. Imputation Techniques for Various Variables 22

Table JAIW.6.2. Seed Values for Multiple Imputation 25

Table JAIW.6.3. Seed Values for Imputation 27

Table JAIW.6.4. Description and Derivation of Primary, Secondary, and Exploratory Efficacy Outcomes 35

Table JAIW.6.5. Description of Primary, Secondary and Exploratory Efficacy Analyses 39

Table JAIW.6.6. Description and Derivation of Health Outcomes and Quality-of-Life Measures 45

Table JAIW.6.7 Description of Health Outcomes and Quality-of-Life Measures Analyses 58

Table JAIW.6.8. Categorical Criteria for Abnormal Treatment-Emergent Blood Pressure and Pulse Measurement, and Categorical Criteria for Weight Changes for Adults 69

Table of Contents

Figure **Page**

Figure JAIW.5.1. Illustration of study design for I4V-MC-JAIW. 13

Figure JAIW.6.1. Illustration of graphical multiple testing procedure with initial α allocation and weights. 29

3. Revision History

Statistical Analysis Plan (SAP) Version 1 was based on Protocol I4V-MC-JAIW(a) and was approved prior to the first unblinding.

Statistical Analysis Plan (SAP) Version 2 was based on Protocol I4V-MC-JAIW(c) and Program Safety Statistical Analysis plan (PSAP) Version 6. It was approved prior to Week 16 Database lock. It included the following revisions:

- changed the primary endpoint to proportion of patients achieving 75% improvement from baseline using the Eczema Area and Severity Index score (EASI75) at Week 16 for baricitinib 2-mg versus placebo
- changed proportion of patients achieving Investigator's Global Assessment (IGA) of 0 or 1 with a ≥2-point improvement at Week 16 from primary to key secondary endpoint
- updated to reflect rolling over of responder completers to extension study I4V-MC-JAIX
- updated illustration of study design
- updated sample size determination
- added modified baseline observation carried forward (mBOCF) missing data imputation method to efficacy endpoints EASI percentage change from baseline, Atopic Dermatitis Sleep Scale (ADSS) Item 2 change from baseline and skin pain Numeric Rating Scale (NRS) change from baseline
- added additional exploratory objective for topical corticosteroids (TCS) use
- added detail analysis on Patient-Oriented Eczema Measure (POEM), Dermatology Life Quality Index (DLQI), Itch NRS, ADSS, Skin Pain NRS, Hospital Anxiety Depression Scale (HADS) in exploratory analyses
- added time to event – IGA (0.1), EASI75 and 4-point improvement in Itch Numeric Rating Scale (NRS) in exploratory analysis
- updated daily diaries to use an interval mean weekly score
- updated or added more derived data: age, body mass index (BMI) category, weight category, cyclosporine inadequate efficacy response, cyclosporine intolerance, cyclosporine contraindication, cyclosporine inadvisable, topical calcineurin inhibitor (TCNI) inadequate efficacy response, TCNI intolerance, TCNI contraindication, TCNI inadvisable
- added baseline value to logistic regression model
- Updatedthe structures of covariance matrix inmixed model repeated measures (MMRM)model
- provided details of efficacy and safety analyses periods
- updated the graphical testing procedure
- updated safety analysis period as the treatment period plus up to 30 days off-drug follow-up time
- updated exposure from weeks to days
- updated Opportunistic Infection section, Skin Infections section, Arterial Thromboembolic (ATE) Events section, Malignancies section and Allergic Reactions/Hypersensitivity section
- removed details included in the standards

4. Study Objectives

4.1. Primary Objective

The primary objective of this study is to test the hypothesis that baricitinib 2-mg once daily (QD) is superior to placebo in the treatment of patients with moderate to severe atopic dermatitis (AD), as assessed by the proportion of patients achieving EASI75 at Week 16.

The associated estimand for this objective is to measure the effect of baricitinib therapy as assessed by the proportion of patients achieving EASI75 at Week 16 assuming the treatment response disappears after patients are rescued or discontinue from study or treatment. See also Sections 6.4 and 6.4.1 on how this estimand handles outcomes after occurrence of any intercurrent event through non-responder imputation (NRI).

4.2. Secondary Objectives

4.2.1. Key Secondary Objectives

These are prespecified objectives that will be adjusted for multiplicity.

Objective	**Endpoint**
To compare the efficacy of baricitinib 1-mg once daily (QD) or 2-mg QD to placebo in atopic dermatitis (AD) during the 16-week double-blind placebo-controlled treatment period as measured by improvements in signs and symptoms of AD.	• Proportion of patients achieving Investigator's Global Assessment (IGA) of 0 or 1 with a ≥2-point improvement at Week 16 • Proportion of patients achieving 75% improvement from baseline using the Eczema Area and Severity Index score (EASI75) at Week 16 (1-mg) • Proportion of patients achieving 90% improvement from baseline using the Eczema Area and Severity Index score (EASI90) at Week 16 • Mean percent change from baseline in EASI score at Week 16 • Proportion of patients achieving SCORing Atopic Dermatitis (SCORAD75) at Week 16
To compare the efficacy of baricitinib 1-mg QD or 2-mg QD to placebo in AD during the 16-week, double-blind, placebo-controlled treatment period as assessed by patient-reported outcome measures	• Proportions of patients achieving a 4-point improvement in Itch Numeric Rating Scale (NRS) at 1 week, 2 weeks, 4 weeks, and 16 weeks • Mean change from baseline in the score of Item 2 of the Atopic Dermatitis Sleep Scale (ADSS) at 1 week and 16 weeks • Mean change from baseline in Skin Pain NRS at Week 16

4.2.2. Other Secondary Objectives

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

Objective	Endpoint
To compare the efficacy of baricitinib 1-mg QD or 2-mg QD to placebo in AD during the 16-week, double-blind, placebo-controlled period as measured by improvement in signs and symptoms of AD	• Proportion of patients achieving Investigator's Global Assessment (IGA) of 0 or 1 with a ≥2-point improvement at Week 4 • Proportion of patients achieving 50% improvement from baseline using the Eczema Area and Severity Index score (EASI50) at Week 16 • Proportion of patients achieving IGA of 0 at Week 16 • Mean change from baseline in SCORing Atopic Dermatitis (SCORAD) at Week 16 • Proportion of patients achieving SCORAD90 at Week 16 • Mean change from baseline in body surface area affected at Week 16 • Proportion of patients developing skin infections requiring antibiotic treatment by Week 16
To compare the efficacy of baricitinib 1-mg QD or 2-mg QD to placebo in AD during the 16-week, double-blind, placebo-controlled treatment period as assessed by patient-reported outcome/quality of life (QoL) measures	• Mean percent change from baseline in Itch Numerical Rating Scale (NRS) at 1 week and 16 weeks • Mean change from baseline in Itch NRS at 4 weeks and 16 weeks • Mean change from baseline in the total score of the Patient-Oriented Eczema Measure (POEM) at Week 16 • Mean change in Patient Global Impression of Severity (PGI-S-AD) scores at Week 16 • Mean change from baseline in the Hospital Anxiety Depression Scale (HADS) at Week 16 • Mean change in Dermatology Life Quality Index (DLQI) scores at Week 16 • Mean change in Work Productivity and Activity Impairment: Atopic Dermatitis (WPAI-AD) scores at Week 16 • Mean change in European Quality of Life–5 Dimensions–5 Levels (EQ-5D-5L) scores at Week 16

4.2.3. Other Secondary Objectives for Week 16 Responders

The secondary objectives and corresponding endpoints for those who responded at Week 16 and continue in the study beyond 16 weeks are as follows:

Objective	Endpoint
To describe the long-term efficacy of baricitinib 1-mg QD or 2-mg QD in AD as measured by improvement in signs and symptoms of AD	• Proportion of patients with a response of Investigator's Global Assessment (IGA) 0 or 1 at Week 16 who maintain an IGA 0 or 1 at Weeks 28, 52, and 104 • Proportion of patients with a response of IGA 0 or 1 at Week 16 who achieve Eczema Area and Severity Index score (EASI75) assessed at Weeks 28, 52, and 104 • Proportion of patients with a response of IGA 0 or 1 at Week 16 who achieve SCORing Atopic Dermatitis (SCORAD75) at Weeks 28, 52, and 104 • Mean percent change from baseline in EASI score at Weeks 28, 52, and 104 • Mean percent change from baseline in SCORAD score at Weeks 28, 52, and 104 • Mean percent change from baseline in SCORAD pruritus at Weeks 28, 52, and 104 • Mean percent change from baseline in Patient-Oriented Eczema Measure (POEM) at Weeks 28, 52, and 104

4.3. Exploratory Objectives

The exploratory objectives of this study are as follows:

Objective
Exploratory objectives evaluating the response to baricitinib treatment regimens on other patient reported outcomes may include dichotomous endpoints or change from baseline for the following measures: Patient-Oriented Eczema Measure (POEM), Dermatology Life Quality Index (DLQI), Itch Numerical Rating Scale (NRS), Atopic Dermatitis Sleep Scale (ADSS), Skin Pain NRS, Hospital Anxiety Depression Scale (HADS), Eczema Area and Severity Index score (EASI) and SCORing Atopic Dermatitis (SCORAD). Topical corticosteroids(TCS) use will be explored. • Time to First Eczema Area and Severity Index score (EASI) 75% Reduction Response • Time to First Investigator's Global Assessment IGA (0,1) Response • Time to First Itch 4 point reduction Response • Proportions of patients achieving a Skin pain 4-pt improvement for those with baseline Skin pain ≥4 • Proportions of patients achieving a DLQI 4-pt improvement for those with baseline DLQI ≥4 • Proportions of patients achieving a DLQI 5 score or less for those with baseline DLQI >5 • Proportions of patients achieving a DLQI 0 or 1 • Proportions of patients achieving a POEM 4-pt improvement for those with baseline POEM ≥4 • Proportions of patients achieving a HADS Anxiety < 8 for those with baseline HADS A ≥8 • Proportions of patients achieving a HADS Depression < 8 for those with baseline HADS D ≥8 • Proportions of patients achieving a HADS A or HADS D < 8 for those with baseline HADS A ≥8 or HADS D ≥8 • Hospital Anxiety Depression Scale total score change from baseline using mixed model repeated measures (MMRM)

Objective
• Proportions of patients achieving a ADSS2 1.5-pt improvement for those with baseline ADSS2 ≥1.5 • Mean change from baseline in the score of Item 1 of the ADSS at 1 week and 16 weeks • Mean change from baseline in the score of Item 3 of the ADSS at 1 week and 16 weeks • Mean change from baseline in EASI score at Week 16 • Mean percentage change from baseline in SCORing Atopic Dermatitis (SCORAD) score at Week 16 • To assess the use of rescue therapy through the weight of sponsor provided topical corticosteroid and number of days not using rescue therapy

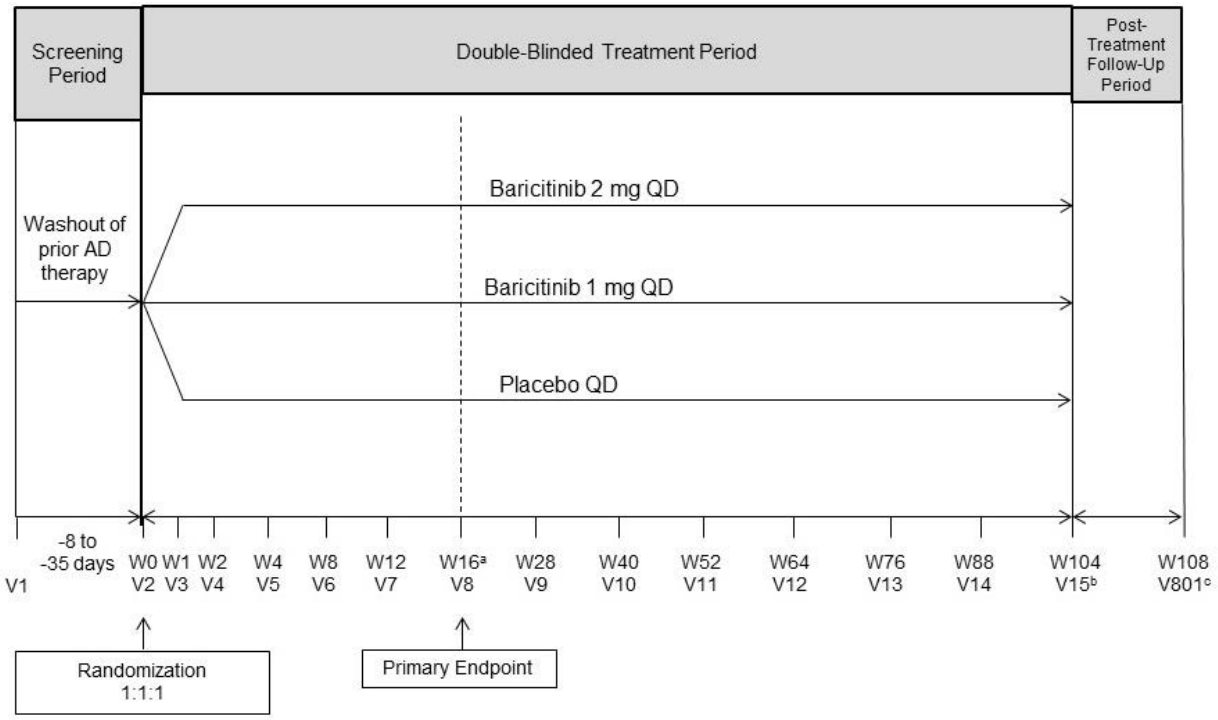
5. Study Design

Study I4V-MC-JAIW (JAIW) is a multicenter, randomized, double-blind, placebo-controlled, parallel-group, outpatient Phase 3 study evaluating the efficacy and safety of baricitinib 1-mg QD and 2-mg QD as compared to placebo in adult patients with moderate to severe AD. The study is divided into 3 periods: a 5-week Screening Period, a 104-week Double-Blinded Treatment Period, and a 4-week Post-Treatment Follow-Up Period.

Approximately 450 patients aged ≥18 years who have responded inadequately to or who are intolerant of topical therapy will be randomized at a 1:1:1 ratio to receive placebo QD, baricitinib 1-mg QD, or baricitinib 2-mg QD (approximately 150 patients per group). Patients will be stratified at randomization according to disease severity (IGA 3 vs. 4).

The primary and key secondary endpoints are assessed prior to or at Visit 8 (Week 16). At Week 16 those patients who met IGA 0 or 1 and who have not required rescue therapy prior to Week 16 will be allowed to continue in this study. All other patients will be discontinued from this study and may be eligible to enroll in a separate open-label study (Study JAIX). Patients experiencing worsening in disease severity resulting in an IGA score of ≥3 after Week 16 of Study JAIW will also have to be discontinued from this study and may be eligible to enroll in the open-label Study JAIX. Patients who complete Week 104 (Visit 15) will have the option to transition to open-label Study JAIX, if eligibility criteria are met regardless of responder status, or continue to the post-treatment follow-up.

Figure JAIW.5.1 illustrates the study design.



Abbreviations: AD = atopic dermatitis; QD = once daily; V = visit; W = week.

[a] At Week 16, all patients who achieve an IGA 0 or 1 and who have not required rescue therapy before Week 16 will be allowed to continue in this study. All other patients will be discontinued from this study and may be eligible to enroll in the separate open-label Study JAIX.

[b] Patients who complete this study will be eligible for assessment to enroll in open-label Study JAIX.

[c] Occurs approximately 28 days after the last dose of investigational product.

Figure JAIW.5.1. Illustration of study design for I4V-MC-JAIW.

5.1. Method of Assignment to Treatment

Patients who meet all criteria for enrollment will be randomized in a 1:1:1 ratio (placebo, baricitinib 1-mg, or baricitinib 2-mg) to double-blind treatment at Visit 2 (Week 0). Randomization will be stratified by disease severity at baseline (IGA 3 or 4).

Assignment to treatment groups will be determined by a computer-generated random sequence using an interactive web-response system (IWRS). The IWRS will be used to assign bottles, each containing double-blind investigational product tablets to each patient, starting at Visit 2 (Week 0) up to and including Visit 14 (Week 88).

6. A Priori Statistical Methods

6.1. Determination of Sample Size

Study JAIW will aim to enroll approximately 450 patients aged ≥18 years. The proposed sample size will ensure at least 68% power to detect any differences between the baricitinib 2-mg and placebo treatment groups, assuming a 10% placebo and 20% baricitinib 2-mg response rate for the primary endpoint EASI75 using a Chi-squared test with a 2-sided α level of 0.05. The assumptions are based on what was observed in the Phase3 monotherapy Studies JAHL and JAHM.

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

6.2. General Considerations

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

Statistical analysis of this study will be the responsibility of Eli Lilly and Company (Lilly). 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 can be made 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 (eg, graphical multiple testing strategy in Section 6.6).

Data collected at early termination visits will be mapped to the closest scheduled visit number for that patient if it falls within the visit window as discussed in Section 6.2.3. 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 to endpoint using modified last observation carried forward (mLOCF) for efficacy analyses, and other categorical analyses including safety.

6.2.1. Reporting Periods

This study will have data locks as described below.

- An unblinded interim lock will be executed after all patients have completed the Week 16 visit (Visit 8) or discontinued. As the primary endpoint (Section 4.2.1) and several secondary endpoints (Section 4.2.2) including safety are evaluated at Week 16, a CSR will report these data.

- An unblinded lock will occur to support the 4-Month Safety Update Report. This lock will include safety, disposition, and exposure data. Additional safety locks to support global submissions may also be performed. Clinical study reports will not be developed based on these safety locks.
- There will be several unblinded locks in support of a Data Monitoring Committee (DMC). Additionally, there will be several blinded locks in support of trial level safety reviews and the Periodic Safety Update Report.
- A final lock will be performed after all patients have completed the Post-Treatment Follow-up Visit (Week 108, Visit 801). A CSR will be developed based on this lock.

The scope of this SAP will be to support the Week 16 unblinding interim lock.

The Blinding/Unblinding Plan for Study JAIW will outline efforts to ensure the blinding integrity after unblinded transfers.

6.2.2. Analysis Populations

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

Per-protocol set (PPS) population: The PPS of the ITT population analysis set will include those patients who do not have any identified important protocol violations considered to impact efficacy analyses. Qualifications for and identification of significant or important protocol violations will be determined while the study remains blinded, prior to database lock.

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 the issue of selection bias. Patients will be analyzed according to the treatment to which they were randomized. In addition, the primary will be repeated using the PPS population.

Efficacy will be summarized in 2 efficacy analysis sets:

- Weeks 0 to 16, and
- after Week 16 to Week 104.

The long-term summary of the JAIW ITT population will include up to 104 weeks. By design all patients continuing past Week 16 of JAIW were responders without rescue mediation. As it is assumed there will be few placebo responders continuing past Week 16, long-term efficacy will be characterized using descriptive statistics after Week 16 up to Week 104 using observed cases.

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 post-baseline visit. This definition excludes patients with no safety assessments post-baseline so that incidence rates are not underestimated.

Safety analyses will be done 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.

The following are the treatment groups for the analysis of safety for Study JAIW for the 16-week interim analyses:

Treatment Group	Definition
Placebo	Placebo at entry to Study JAIW followed up to data cut (interim clinical study report [CSR])
Baricitinib 1-mg	Baricitinib 1-mg at entry to Study JAIW followed up to data cut (interim CSR)
Baricitinib 2-mg	Baricitinib 2-mg at entry to Study JAIW followed up to data cut (interim CSR)

In the rare situation where a patient is lost to follow-up at the first post-baseline visit, but some safety data exists (eg, unscheduled laboratory assessments) after first dose of study drug, a listing of the data or a patient profile will be provided, when requested.

For the unblinded interim lock which is executed after all patients have completed the Week 16 visit (Visit 8) or discontinued, the efficacy analysis will include up to Week 16 and the safety analysis will be up to Week 16 and selected analyses up to the data cut date.

6.2.3. Definition of Baseline and Post-Baseline 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 defined as the mean of the non-missing assessments in the 7 days prior to the date of first study drug administration.

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, 1 day at a time, to obtain the 4 most recent 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.

Post-Baseline

Post-baseline measurements are collected after study drug administration for electronic patient-reported outcome (ePRO), Itch NRS, Skin Pain NRS, ADSS and PGI-S-AD up to Week 16 (Visit 8). Other post-baseline measurements are collected up to Week 104 (Visit 15).

Non-missing efficacy data collected at scheduled visits (e.g., = Electronic version of Clinical Outcome Assessment [eCOA], clinician-reported outcome [ClinRO]) will be used for analyses.

If an assessment is missing at a scheduled visit, an unscheduled post-baseline assessment can be used provided it falls within a ±4 day window of the scheduled visit date. 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 2 unscheduled visits of equal distance are available, then the latter of the 2 will be used. If there is no non-missing measure collected at the scheduled visit, or an unscheduled visit falling within the visit window, the assessment is missing for that scheduled visit.

For Treatment Period 1, post-baseline daily diary endpoints will be the mean of weekly visit windows (diary windows) anchored on day of first dose (Day 1) for Week 1 to Week 14 as follows:

Week	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Day	1-7	8-14	15-21	22-28	29-35	36-42	43-49	50-56	57-63	64-70	71-77	78-84	85-91	92-98

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 3.

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 and up to Visit 8 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 or 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 Day 106 to Day 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 Week 15 and Week 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 1 diary record on a particular date, the first record on that particular date will be used in the analysis.

Post-baseline 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.

6.2.4. Analysis Methods

Unless otherwise specified, all analyses described in this section will compare estimates (eg, odds ratios, least square means, proportions) of baricitinib 1-mg and 2-mg to placebo. Thus odds ratios are baricitinib treatment groups relative to placebo; similarly, least-square mean (LSM) differences and differences in proportions are between baricitinib treatment groups and placebo.

The main analysis method of categorical efficacy variables and health outcomes variables will use a logistic regression analysis with baseline disease severity (IGA), baseline value, and treatment group in the model. 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, 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 LSMs 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 100(1-alpha) CI) will be used for statistical inference. The LSM difference, standard error, p-value and 100(1-alpha)% CI will be reported.

Treatment comparisons for continuous efficacy and health outcomes variables may also be made using analysis of covariance (ANCOVA) for key secondary and secondary objectives. When an ANCOVA model is used, the model includes-baseline disease severity (IGA), treatment group, and baseline value. Treatment LSMs 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 100(1-alpha) CI will be used for statistical inference.

Fisher's exact test will be used to test the difference between each baricitinib dose and placebo in proportion of patients experiencing adverse events (AEs), discontinuation from study drug, and 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.

Time to event will be analyzed using cumulative incidence function with observed values, defining first time reaching the event IGA (0,1), EASI 75, or Itch NRS 4-pt improvement before rescue as onset, treating rescue and discontinuing for lack of efficacy as competing event censor up to Week 16.

6.2.5. Derived Data

The following endpoints are derived from collected data:

- Age (year), derived using first dose date as the reference start date and July 1 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 consent 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 (year) = [(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: (number of months between date of AD diagnosis and July 1 of birth year) / 12, and truncated to a whole-integer age
- Diagnosis age group (<18, ≥18 and <50, ≥50 years old)
- Change from baseline = post-baseline 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 response(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)
  - 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:
    - Reason for not using medication: contraindication, unfavorable benefit/risk, or physician decision
    - Reason for discontinuation: inadequate response, intolerance to medication, or contraindication.
- topical calcineurin inhibitor inadequate efficacy response (yes, no)
  - Set **yes** if the reason for discontinuation is inadequate response
- topical calcineurin inhibitor intolerance (yes, no)
  - Set **yes** if the reasons for discontinuation are: intolerance to medication or contraindication (Physician indicated TCNI was used and a contraindication was noted)
- topical calcineurin inhibitor contraindication / [ineligible](yes, no)
  - Set to **yes** if TCNI never used because of a contraindication
- topical calcineurin inhibitor inadvisable (yes, no)

- Set to **yes** if the following reasons were selected for either not using the medication or discontinuing the medication:
  - Reason for not using medication: Physician decision, concern about side effects, unfavorable benefit risk, contraindication
  - Reasons for discontinuation: inadequate response, intolerance to medication, or contraindication

6.3. Adjustments for Covariates

The randomization to treatment groups at Week 0 (Visit 2) is stratified by disease severity (IGA) as described in Section 5.1. Unless otherwise specified, the statistical analysis models will control for disease severity. The covariates used in the logistic model for categorical data will include the parameter value at baseline. The covariates used in the ANCOVA models for continuous data generally will include the parameter's value at baseline. Inclusion of baseline in the ANCOVA models ensures treatment LSMs 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 (International Conference on Harmonisation [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, after rescue therapy)
- discontinuation from the study
- missing an intermediate visit prior to discontinuation or rescue, and
- lost to follow-up.

Non-censor intercurrent events are events that are not due to the application of any censoring rule (ie, 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. Alternatively, 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. As patients who are rescued to systemic therapies are required to permanently discontinue study drug, they will also have post-rescue observations censored. The secondary censoring rule will be applied to primary and key secondary efficacy and health outcome endpoints.

Non-responder imputation (for categorical variables) and MMRM (for continuous variables) will be the primary methods used to handle missing data. Censoring rules, along with their associated estimator assumptions, are described in Sections 6.4.1 through 6.4.6.

Table JAIW.6.1 summarizes how imputation techniques and censoring rules are applied to efficacy and health outcome endpoints.

Table JAIW.6.1. Imputation Techniques for Various Variables

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

Abbreviations: AD = atopic dermatitis; ADSS = Atopic Dermatitis Sleep Scale; EASI = Eczema Area and Severity Index score; IGA = Investigator's Global Assessment for AD; mBOCF = baseline observation carried forward; 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. Non-Responder 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. For analyses which utilize either of the censoring methods, randomized patients without at least 1 post-baseline observation will also be defined as non-responders for all visits. In addition, patients who are missing a value prior to discontinuation or

rescue (if censoring on rescue) (ie, the patient is missing an intermediate visit) will be imputed as non-responders at that visit.

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 imputation 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 key secondary continuous endpoints will also utilize MMRM after applying the secondary censoring rule (Table JAIW.6.1).

6.4.3. Modified Last Observation Carried Forward

For continuous measure, a 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 intercurent 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 post-baseline 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 key secondary and secondary endpoints will use mLOCF imputation methodology with an ANCOVA as sensitivity analyses to the MMRM analyses.

6.4.4. Modified Baseline Observation Carried Forward

A baseline observation analysis is performed by carrying forward the baseline assessment for the continuous measures, assuming that effect of treatment will loss and patient status will return to the baseline status after the occurrence of the intercurrent event (after application of the primary censoring rule). After mBOCF imputation, data from patients with nonmissing baseline will be included in the analyses. These mBOCF analyses will be applied to ITT population and on key secondary continuous efficacy and health outcomes endpoints.

6.4.5. Placebo Multiple Imputation

The Placebo Multiple Imputation (pMI) methodology will be used as a sensitivity analysis for the analysis of the EASI75 efficacy endpoint 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 are 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 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.

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 of EASI75, IGA(0,1), 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 into the analysis model. A logistic regression model will be applied. For continuous endpoints, fit its treatment response model (MMRM) for each completed data set. For the primary and secondary key efficacy endpoints of EASI75, IGA(0,1), 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 into the analysis model. A logistic regression model will be applied.

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 JAIW.6.2.

Table JAIW.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 weeks 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: AD = atopic dermatitis; 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.

6.4.6. Tipping Point Analyses

To investigate the missing data mechanism, additional analyses using multiple imputation (MI) under the missing not at random assumption will be provided for the following primary and key secondary objectives:

- EASI75 at Week 16, baricitinib 2-mg compared to placebo
- IGA (0,1) with ≥2-point improvement at Week 16, baricitinib 2-mg compared to placebo
- Itch NRS 4-point improvement from baseline to Week 16, baricitinib 2-mg compared to placebo

All patients in the ITT population will be included in the analyses. 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 1-mg or 2-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 (Section 6.2.4).

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 (MCMC) 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 6.2.4). Results across the imputed datasets are aggregated using SAS® Proc MI ANALYZE 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-dimensional 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-dimensional 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, 0.05, … 0.3, as appropriate for the data.
2. For missing responses in the placebo group, a range of responses probabilities (eg, probability = 0, 0.05 … 0.3 by incremental 0.05, increments may be changed after unblinding to best reflect reasonable response rates, as appropriate for the data) 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 6.2.4). 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 (eg, all missing responses in the placebo and baricitinib groups are imputed as responders and non-responders, 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. 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 JAIW.6.3.

Table JAIW.6.3. Seed Values for Imputation

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

Abbreviations: EASI = Eczema Area and Severity Index score; IGA = Investigator's Global Assessment for atopic dermatitis; NRS = Numeric Rating Scale.

6.5. Multicenter Studies

This study will be conducted by multiple investigators at multiple sites in the US and Canada. Country differences will be assessed in the subgroup analysis as discussed in Section 6.15.

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. The subscript for **H** denotes dose (2-mg, 1-mg), the numerical identifier of the endpoint within the dose, and the type of hypothesis (0 for null, 1 for alternative), respectively.

Primary Null Hypotheses:

- . , , : Proportion of baricitinib 2-mg patients achieving EASI75 is less than or equal to the proportion of placebo patients achieving EASI75 at Week 16

Key Secondary Null Hypotheses:

- $\mathbf{H_{2,2,0}}$: Mean percent change from baseline in EASI score for baricitinib 2-mg patients is less than or equal to the percent change from baseline in EASI score for placebo patients at Week 16
- $\mathbf{H_{2,3,0}}$: Proportion of baricitinib 2-mg patients achieving a 4-point improvement in Itch NRS is less than or equal to the proportion of placebo patients achieving a 4-point improvement in Itch NRS at Week 16
- $\mathbf{H_{2.4,0}}$: Mean change from baseline in Skin Pain NRS for baricitinib 2-mg patients is less than or equal to the mean change from baseline in Skin Pain NRS for placebo patients at Week 16
- $\mathbf{H_{2,5,0}}$: Proportion of baricitinib 2-mg patients achieving a 4-point improvement in Itch NRS is less than or equal to the proportion of placebo patients achieving a 4-point improvement in Itch NRS at Week 4

- $\mathbf{H_{2,6,0}}$: Proportion of baricitinib 2-mg patients achieving IGA of 0 or 1 with a ≥2-point improvement from baseline at Week 16 is less than or equal to the proportion of placebo patients achieving IGA of 0 or 1 with a ≥2-point improvement from baseline at Week 16 [IGA0-1]
- $\mathbf{H_{2,7,0}}$: Proportion of baricitinib 2-mg patients achieving a 4-point improvement in Itch NRS is less than or equal to the proportion of placebo patients achieving a 4-point improvement in Itch NRS at Week 2
- $\mathbf{H_{2,8,0}}$: Proportion of baricitinib 2-mg patients achieving a 4-point improvement in Itch NRS is less than or equal to the proportion of placebo patients achieving a 4-point improvement in Itch NRS at Week 1
- $\mathbf{H_{2,9,0}}$: Mean change from baseline in the score of Item 2 of the ADSS for baricitinib 2-mg patients is less than or equal to the mean change from baseline in the score of Item 2 of the ADSS for placebo patients at Week 16
- $\mathbf{H_{2,10,0}}$: Mean change from baseline in the score of Item 2 of the ADSS for baricitinib 2-mg patients is less than or equal to the mean change from baseline in the score of Item 2 of the ADSS for placebo patients at Week 1
- $\mathbf{H_{2,11,0}}$: Proportion of baricitinib 2-mg patients achieving EASI90 is less than or equal to the proportion of placebo patients achieving EASI90 at Week 16
- $\mathbf{H_{2,12,0}}$: Proportion of baricitinib 2-mg patients achieving SCORAD75 is less than or equal to the proportion of placebo patients achieving SCORAD75 at Week 16
- $\mathbf{H_{1,1,0}}$: Proportion of baricitinib 1-mg patients achieving EASI75 is less than or equal to the proportion of placebo patients achieving EASI75 at Week 16
- $\mathbf{H_{1,2,0}}$: Percent change from baseline in EASI score for baricitinib 1-mg patients is less than or equal to the percent change from baseline in EASI score for placebo patients at Week 16
- $\mathbf{H_{1,3,0}}$: Proportion of baricitinib 1-mg patients achieving a 4-point improvement in Itch NRS is less than or equal to the proportion of placebo patients achieving a 4-point improvement in Itch NRS at Week 16
- $\mathbf{H_{1.4,0}}$: Mean change from baseline in Skin Pain NRS for baricitinib 1-mg patients is less than or equal to the mean change from baseline in Skin Pain NRS for placebo patients at Week 16
- $\mathbf{H_{1,5,0}}$: Proportion of baricitinib 1-mg patients achieving a 4-point improvement in Itch NRS is less than or equal to the proportion of placebo patients achieving a 4-point improvement in Itch NRS at Week 4
- $\mathbf{H_{1,6,0}}$: Proportion of baricitinib 1-mg patients achieving IGA of 0 or 1 with a ≥2-point improvement from baseline at Week 16 is less than or equal to the proportion of placebo patients achieving IGA of 0 or 1 with a ≥2-point improvement from baseline at Week 16
- $\mathbf{H_{1,7,0}}$: Proportion of baricitinib 1-mg patients achieving a 4-point improvement in Itch NRS is less than or equal to the proportion of placebo patients achieving a 4-point improvement in Itch NRS at Week 2
- $\mathbf{H_{1,8,0}}$: Proportion of baricitinib 1-mg patients achieving a 4-point improvement in Itch NRS is less than or equal to the proportion of placebo patients achieving a 4-point improvement in Itch NRS at Week 1

- **$H_{1,9,0}$**: Mean change from baseline in the score of Item 2 of the ADSS for baricitinib 1-mg patients is less than or equal to the mean change from baseline in the score of Item 2 of the ADSS for placebo patients at Week 16
- **$H_{1,10,0}$**: Mean change from baseline in the score of Item 2 of the ADSS for baricitinib 1-mg patients is less than or equal to the mean change from baseline in the score of Item 2 of the ADSS for placebo patients at Week 1
- **$H_{1,11,0}$**: Proportion of baricitinib 1-mg patients achieving EASI90 is less than or equal to the proportion of placebo patients achieving EASI90 at Week 16
- **$H_{1,12,0}$**: Proportion of baricitinib 1-mg patients achieving SCORAD75 is less than or equal to the proportion of placebo patients achieving SCORAD75 at Week 16

The primary null hypothesis includes testing whether the baricitinib 2-mg is superior to placebo at the primary endpoint, defined as the proportion of placebo patients achieving EASI75 at Week 16. The graphical multiple testing procedure described in Bretz et al. (2011), which is a closed testing procedure will be used; hence, it strongly controls the family-wise error rate across all endpoints (Alosh et al. 2014).

Figure JAIW.6.1 depicts the graphical testing scheme (including testing order, interrelationships, Type I error allocation, and the associated propagation).

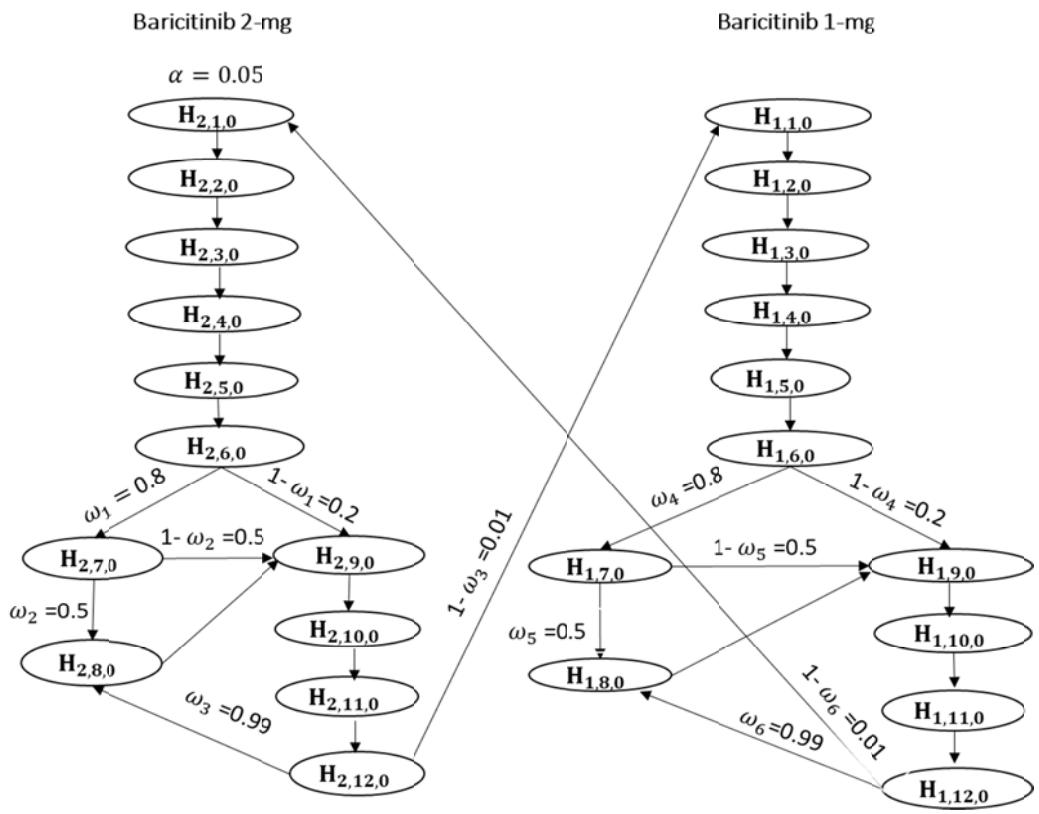


Figure JAIW.6.1. Illustration of graphical multiple testing procedure with initial α allocation and weights.

If $H_{2,1,0}$ is not rejected, no further testing is conducted as the α for that test is considered "spent" and cannot be passed to other endpoints. If $H_{2,1,0}$ is rejected, then α will be propagated to $H_{2,2,0}$. The testing process continues with α propagated according to the weights on the corresponding edges displayed in Figure JAIW.6.1, as long as each hypothesis in the sequence can be rejected at its allocated α level. Each time a hypothesis is rejected, the graph is updated to reflect the reallocation of α, which is considered "recycled" by Alosh et al. (2014). This iterative process of updating the graph and reallocating α is repeated until all hypotheses have been tested or when no remaining hypotheses can be rejected at their corresponding α levels.

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 will be summarized using the ITT population. Frequency counts and percentages of patients will be summarized by treatment group by the following dispositions:

- completed at least the Week 16
- completed the Week 104
- discontinued early from the study
- enrolled in Study JAIX after discontinuation
- rescued
- non-rescued
- reason for discontinuation

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

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)

- country
- weight (kg)
- weight category (<60 kg, ≥60 to <100 kg, ≥100 kg
- height (cm)
- body mass index (kg/m^2)
- body mass index 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 below baseline disease information (although not inclusive) 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 IGA for AD score
- Eczema Area and Severity Index (EASI) score
- SCORing Atopic Dermatitis (SCORAD)
- Body Surface Area affected by AD
- Hospital Anxiety Depression Scale (HADS) subscales
- Patient-Oriented Eczema Measure (POEM)
- Itch 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 topical calcineurin inhibitors (yes, no)
- topical calcineurin inhibitor inadequate response (yes, no)
- topical calcineurin inhibitor intolerance (yes, no)
- topical calcineurin inhibitor contraindication [ineligible] (yes, no)

- topical calcineurin inhibitor inadvisable (yes, no)
- vaccine:
  - zoster vaccine (Yes, No)
  - tuberculosis (TB) vaccine (Yes, No)
- baseline renal function status: impaired (estimated glomerular filtration rate [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.

Pre-existing conditions are defined as those conditions recorded in the Pre-existing Conditions and Medical History eCRF, or the Prespecified Medical History: Comorbidities eCRF with a start date and time prior to the informed consent and with a stop date that is after the informed consent date or have no stop date (ongoing). Adverse events are recorded in the eCRFs. For events recorded on AE page, we considered it as a pre-existing event if it's onset date was before first dose date. 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 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 at each visit using the ITT population.

All patients are expected to take 2 tablets daily as described in the protocol. 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 the dose was withheld will be appropriately adjusted in the 'expected number of total tablets' element of the compliance calculation given below.

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

$$\text{Compliance} = 100 * \frac{\text{total number of tablets dispensed} - \text{total number of tablets returned}}{\text{expected number of total tablets}}$$

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 visit – 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 rates 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.

6.9.1. Rescue Treatment

Descriptive statistics for drug accountability of topical low and moderate potency rescue medication provided by the sponsor will also be supplied, including the amount utilized throughout treatment (from Week 0 through Week 16). The total amount in grams for low and moderate potency will be summarized between scheduled 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), as well as throughout the treatment period from Week 0 through Week 16. The total amount will also be presented for the all visit intervals, irrespective of potency. If a returned tube is not weighed in grams, then the tube can be classified as partially used, fully used, unused, or unknown. Partially used rescue medication tubes will be considered to be 50% used, whereas Fully Used and Unused will be considered as 100% used 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 number of days rescue therapy is used for each patient is also collected on the diary device. The proportion of time that the patients did not use rescue therapy will be summarized for the aforementioned visit intervals by potency (low or moderate) and both potencies combined. For this analysis, the date of the first entry on the diary device will be used to signify the first day of rescue therapy use.

Additionally, a summary of the initial rescue therapy and the reason for rescue will be produced, as well as a summary of the proportion of patients rescued at each study visit up to week 16. A summary of all rescue medications will be provided.

6.10. 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. A summary of previous medications used for AD, as well as zoster and TB vaccine, and medications that are discontinued after screening and before the first dose of study drug, will be prepared using frequency counts and percentages by preferred medication name, with preferred medication names sorted by frequency in the baricitinib 2-mg group. Concomitant therapy will be recorded at each visit and will be classified similarly.

Concomitant therapy for the treatment period is defined as therapy that starts before or during the treatment periods 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 (eg, 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 provided for the following category: previous AD therapies

Summaries of concomitant medications will be provided for the following category: concomitant medications excluding rescue medicine

6.11. Efficacy Analyses

The general methods used to summarize efficacy data, including the definition of baseline value for assessments, are described in Section 6.2. Efficacy analyses will generally be analyzed according to the following formats and patients will be analyzed according to the investigational product to which they were randomized at Week 0 (Visit 2):

- Week 0 to Week 16, with primary censoring rule
- Week 0 to Week 16, with secondary censoring rule for primary and key secondary objectives.
- Week 16 to Week 104, as observed, (Note: will not be completed for the Week-16 primary outcome DBL, but will be completed for subsequent DBLs)

Table JAIW.6.4 includes the descriptions and derivations of the primary, secondary, and exploratory efficacy outcomes.

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

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

Measure	Description	Variable	Derivation/Comment	Imputation Approach if Missing Components
Eczema Area and Severity Index (EASI)	The EASI assesses objective physician estimates of 2 dimensions of AD, 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 0-3 (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 0-72, and the final EASI score will be obtained by weight averaging these 4 scores. Hence, the final EASI score will range from 0-72 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 0-3 and value from percentage involvement is on a scale of 0-6. 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.
		▪ 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
		▪ Change from baseline in EASI score ▪ Percent change from baseline EASI	Change from baseline: observed EASI score – baseline EASI score % change from baseline: $100 \times \frac{Observed\ score - Baseline}{Baseline}$	Missing if baseline or observed value is missing

Measure	Description	Variable	Derivation/Comment	Imputation Approach if Missing Components
		score		
		▪ Time to reaching EASI 75	first time reaching EASI75 as event 1, rescue and discontinue for lack of efficacy as event 2, censor up to Week 16	Use observed value, rescue and discontinue for lack of efficacy as competing event, censor up to Week 16
Body Surface Area (BSA) Affected by AD	BSA 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 (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.	▪ BSA score	Use the percentage of skin affected for each region (0 to 100%) in EASI as follows:	N/A – partial assessments cannot be saved.
		▪ Change from baseline in BSA score	Change from baseline: observed BSA score – baseline BSA score	Missing if baseline or observed value is missing.

Measure	Description	Variable	Derivation/Comment	Imputation Approach if Missing Components
Validated Investigator's Global Assessment for AD (IGA)	The validated IGA 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.
		▪ Time to IGA (0,1)	▪ first time reaching IGA (0,1) as event 1, rescue and discontinue for lack of efficacy as event 2, censor up to Week 16	▪ Use observed value, rescue and discontinue for lack of efficacy as competing event, censor up to Week 16
SCORing Atopic Dermatitis (SCORAD)	The 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	▪ 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: $100 \times \frac{Observed\ score - Baseline}{Baseline}$	Missing if baseline or observed value is missing

Measure	Description	Variable	Derivation/Comment	Imputation Approach if Missing Components
	disease, disease severity, and subjective symptoms combine to give a maximum possible score of 103 (ETFAD 1993; Kunz et al. 1997; Schram et al. 2012).	▪ 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

Abbreviations: AD = atopic dermatitis; N/A = not applicable.

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

Measure	Variable	Analysis Method (Section 6.2.4)	Population (Section 6.2.2)	Comparison[a]/Time Point	Analysis Type
Eczema Area and Severity Index (EASI) [*categorical*]	Proportion of patients achieving EASI75	Logistic regression using NRI (both censoring rules, respectively)	ITT	Bari 2 mg or Bari 1 mg vs PBO; Week 16	Primary analysis(Bari 2-mg vs PBO) Key seconday analysis (Bari 1-mg vs PBO) /secondary censoring rule, sensitivity analysis
		Logistic regression using NRI (primary censoring rule)	PPS	Bari 2 mg vs PBO; Week 16	Sensitivity analysis
		Logistic regression using pMI (primary censoring rule)	ITT	Bari 2 mg or Bari 1 mg vs PBO; Week 16	Sensitivity analysis
		Logistic regression using Tipping Point (primary censoring rule)	ITT	Bari 2 mg or Bari 1 mg vs PBO; Week 16	Sensitivity analysis
		Cumulative Incidence Function of Time to reaching EASI 75% reduction (primary censoring rule)	ITT	Bari 2 mg or Bari 1 mg vs PBO; Week 16	Exploratory analysis
	Proportion of patients achieving EASI90	Logistic regression using NRI (both censoring rules, respectively)	ITT	Bari 2 mg or Bari 1 mg vs PBO; Week 16	Key secondary analysis/secondary censoring rule, sensitivity analysis
		Logistic regression using pMI (primary censoring rule)	ITT	Bari 2 mg or Bari 1 mg vs PBO; Week 16	Sensitivity analysis
	Proportion of patients achieving EASI50	Logistic regression using NRI (primary censoring rule)	ITT	Bari 2 mg or Bari 1 mg vs PBO; Week 16	Secondary analysis
Eczema Area and Severity Index (EASI) [*continuous*]	EASI score % change from baseline	MMRM (both censoring rules, respectively)	ITT	Bari 2 mg or Bari 1 mg vs PBO; Week 16	Key secondary analysis/secondary censoring rule, sensitivity analysis
		ANCOVA; pMI (primary censoring rule)	ITT	Bari 2 mg or Bari 1 mg vs PBO; Week 16	Sensitivity analysis

Measure	Variable	Analysis Method (Section 6.2.4)	Population (Section 6.2.2)	Comparison[a]/Time Point	Analysis Type
		ANCOVA; mLOCF (primary censoring rule)	ITT	Bari 2 mg or Bari 1 mg vs PBO; Week 16	Sensitivity analysis
		ANCOVA; mBOCF (primary censoring rule)	ITT	Bari 2 mg or Bari 1 mg vs PBO; Week 16	Sensitivity analysis
	EASI score change from baseline	MMRM (primary censoring rule)	ITT	Bari 2 mg or Bari 1 mg vs PBO; Week 16	Exploratory analysis
Validated Investigator's Global Assessment for AD (IGA)	Proportion of patients achieving IGA [0,1] with a ≥2-point improvement	Logistic regression using NRI (both censoring rules, respectively)	ITT	Bari 2 mg or Bari 1 mg vs PBO; Week 16	Key secondary analysis/secondary censoring rule, sensitivity analysis
		Logistic regression using pMI (primary censoring rule)	ITT	Bari 2 mg or Bari 1 mg vs PBO; Week 16	Sensitivity analysis
		Tipping point analysis (primary censoring rule)	ITT	Bari 2 mg or Bari 1 mg vs PBO; Week 16	Sensitivity analysis
		Cumulative Incidence Function of Time to reaching IGA (0,1) reduction (primary censoring rule)	ITT	Bari 2 mg, 1 mg vs PBO; Weeks 0 to 16	Exploratory analysis
	Proportion of patients achieving IGA [0,1] with a ≥2-point improvement	Logistic regression using NRI (primary censoring rule)	ITT	Bari 2 mg or Bari 1 mg vs PBO; Week 4	Secondary analysis
	Proportion of patients achieving IGA [0]	Logistic regression using NRI (primary censoring rule)	ITT	Bari 2 mg or Bari 1 mg vs PBO; Week 16	Secondary analysis
Body Surface Area (BSA) Affected by AD	BSA change from baseline	MMRM & ANCOVA; mLOCF (primary censoring rule)	ITT	Bari 2 mg or Bari 1 mg vs PBO; Week 16	Secondary analysis/ANCOVA with mLCOF, sensitivity analysis

Measure	Variable	Analysis Method (Section 6.2.4)	Population (Section 6.2.2)	Comparison[a]/Time Point	Analysis Type
SCORing AD (SCORAD) [*categorical*]	Proportion of patients achieving SCORAD75	Logistic regression using NRI (both censoring rules, respectively)	ITT	Bari 2 mg or Bari 1 mg vs PBO; Week 16	Key secondary analysis/secondary censoring rule, sensitivity analysis
		Logistic regression using pMI (primary censoring rule)	ITT	Bari 2 mg or Bari 1 mg vs PBO; Week 16	Sensitivity analysis
	Proportion of patients achieving SCORAD90	Logistic regression using NRI (primary censoring rule)	ITT	Bari 2 mg or Bari 1 mg vs PBO; Week 16	Secondary analysis
SCORing AD (SCORAD) [*continuous*]	SCORAD score change from baseline	MMRM & ANCOVA; mLOCF (primary censoring rule)	ITT	Bari 2 mg or Bari 1 mg vs PBO; Week 16	Secondary analysis/ANCOVA with mLCOF, sensitivity analysis
	SCORAD score % change from baseline	MMRM (primary censoring rule)	ITT	Bari 2 mg or Bari 1 mg vs PBO; Week 16	Exploratory analysis
Skin Infections	Proportion of patients developing skin infections requiring antibiotic treatment	Fisher's exact	ITT	Bari 2 mg or Bari 1 mg vs PBO; Week 16	Secondary analysis

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

6.11.1. Primary Outcome and Methodology

Both EASI score and IGAs are commonly used in clinical trials, both for qualifying patients for enrollment and for evaluating treatment efficacy (Langley et al. 2015; Futamura et al. 2016; Bożek and Reich, 2017). There is no single "gold standard" disease severity scale for AD; however, IGA scales provide clinically meaningful measures to patients and investigators that are easily described and that correspond to disease severity categories (for example, moderate to severe), and a 75% improvement from Baseline (EASI75) is a commonly used measure of treatment effect in AD clinical trials.

The primary analysis of the study is to test the hypothesis that baricitinib 2-mg is superior to placebo when evaluating the proportion of patients achieving EASI75 at Week 16 using the ITT population, assuming the treatment response disappears after patients are rescued or discontinue from study or 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 hypothesis that baricitinib 2-mg is superior to placebo when evaluating the proportion of patients achieving EASI75 at Week 16 using the ITT population, assuming the treatment response disappears after patients discontinue from study or 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 will be used for the treatment group comparisons. The odds ratio, corresponding 95% CIs, and p-value, as well as the treatment differences and the corresponding 95% CIs, will be reported. Missing data will be imputed using the NRI method described in Section 6.4.1.

Multiplicity-controlled analyses will be performed on the primary and key secondary (see Section 4.2.1) objectives in order 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.

6.11.2. Secondary and Exploratory Efficacy Analyses

For secondary analysis, the null hypotheses is that neither baricitinib 2-mg nor baricitinib 1-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 JAIW.6.1.

There will be no adjustment for multiple comparisons for any other analyses. The secondary and exploratory efficacy analyses are detailed in Table JAIW.6.5. Health outcomes analyses are described in Section 6.12.

6.11.3. 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 6.2.2)
- Analyses of key endpoints using the secondary censoring rule (Section 6.4)
- Placebo multiple imputation (Section 6.4.5)
- Tipping point analysis (Section 6.4.6)
- Analysis of continuous outcomes with ANCOVA (Section 6.2.4), with missing data imputed using mLOCF (Section 6.4.3).
- Analysis of continuous outcomes with ANCOVA (Section 6.2.4), with missing data imputed using mBOCF (Section 6.4.4).

6.12. Health Outcomes/Quality-of-Life Analyses

The general methods used to summarize health outcomes and QoL measures, including the definition of baseline value for assessments, are described in Section 6.2.

Health outcomes and QoL measures will generally be analyzed according to the formats discussed in Section 6.11.

Table JAIW.6.6 includes the descriptions and derivations of the health outcomes and QoL measures.

Table JAIW.6.7 provides the detailed analyses, including analysis type, method and imputation, population, time point, and comparisons for health outcomes and QoL measures.

Additional psychometric analyses will be performed by Global Patient Outcomes Real World Evidence at Lilly and documented in a separate analysis plan.

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

Measure	Description	Variable	Derivation/Comment	Imputation Approach if with Missing Components
Itch Numeric Rating Scale (NRS)	The Itch 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 6.2.3 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 6.2.3 on how to derive the visit score.	Refer to Section 6.2.3 on how to derive the 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: $100 \times \frac{Observed\ score - Baseline}{Baseline}$	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
		▪ Cumulative Incidence Function of Time to reaching Itch NRS 4-pt improvement (primary censoring rule)	first time reaching Itch NRS 4-pt improvement as event 1, rescue and discontinue for lack of efficacy as event 2, censor up to week 16	Use observed value, rescue and discontinue for lack of efficacy as competing event, censor up to week 16

Description and Derivation of Health Outcomes and Quality-of-Life Measures

Measure	Description	Variable	Derivation/Comment	Imputation Approach if with Missing Components
Skin Pain NRS	The 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 6.2.3 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- 10. Refer to Section 6.2.3 on how to derive the visit score.	Refer to Section 6.2.3 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
		▪ 4-point Skin Pain improvement in subgroup of patients with baseline Skin Pain NRS ≥4	Change from baseline ≤-4 and baseline ≥4	Missing if baseline is missing or <4 or observed value is missing

Description and Derivation of Health Outcomes and Quality-of-Life Measures

Measure	Description	Variable	Derivation/Comment	Imputation Approach if with Missing Components
Atopic Dermatitis Sleep Scale (ADSS)	The 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. Refer to Section 6.2.3 for details on how to calculate the weekly score which will be used in the continuous analysis.	▪ 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 6.2.3 on how to derive the visit score.	Refer to Section 6.2.3 on how to derive the 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.
		▪ 1.5 point improvement on Item 2 of ADSS	Change from baseline <= -1.5 and baseline >=1.5 in score of Item 2 of ADSS	Missing if baseline is missing or <1.5 or observed value is missing.

Description and Derivation of Health Outcomes and Quality-of-Life Measures

Measure	Description	Variable	Derivation/Comment	Imputation Approach if with 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 POEM improvement in subgroup of patients with baseline POEM score ≥4	Change from baseline ≤ -4 and baseline ≥4	Missing if baseline is missing or <4 or observed value is missing.
Patient Global Impression of Severity–Atopic Dermatitis (PGI-S-AD)	The 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." Refer to Section 6.2.3 for details on how to calculate the weekly score which will be used in the continuous analysis.	PGI-S-AD score	Single item. Range 1 to 5. Refer to Section 6.2.3 on how to derive the visit score.	Refer to Section 6.2.3 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.

Description and Derivation of Health Outcomes and Quality-of-Life Measures

Measure	Description	Variable	Derivation/Comment	Imputation Approach if with Missing Components
Hospital Anxiety Depression Scale (HADS)	The 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 (eg, 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 domain	Change from baseline: observed HADS domain score – baseline HADS domain score	Missing if baseline or observed value is missing.
		Change from baseline in HADS total	Change from baseline: observed HADS domain score – baseline HADS total score	Missing if baseline or observed value is missing.
		▪ HADS Anxiety <8 in subgroup of patients with baseline HADS Anxiety score ≥8 ▪ HADS Depression <8 in subgroup of patients with baseline HADS Depression score ≥8	observed HADS post-baseline <8 and baseline score >=8 for each HADS domain score	Missing if baseline is missing or <8 or observed value is missing.

Description and Derivation of Health Outcomes and Quality-of-Life Measures

Measure	Description	Variable	Derivation/Comment	Imputation Approach if with Missing Components
Hospital Anxiety Depression Scale (HADS)	The 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 (eg, 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 Anxiety or Depression score <8 in subgroup of patients with baseline HADS Anxiety or Depression score ≥8	observed HADS post-baseline <8 and baseline score >=8 for any HADS domain score	Missing if baseline is missing or <8 or observed value is missing for both domain

Description and Derivation of Health Outcomes and Quality-of-Life Measures

Measure	Description	Variable	Derivation / Comment	Imputation Approach if with Missing Components
Dermatology Life Quality Index (DLQI)	The 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 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).	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.
		Work and school domain	Sum of Questions 7 and 7B (if 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 6 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.

Description and Derivation of Health Outcomes and Quality-of-Life Measures

Measure	Description	Variable	Derivation / Comment	Imputation Approach if with Missing Components
Dermatology Life Quality Index (DLQI)	The 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 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).	DLQI total score <=5 in subgroup of patients who had baseline DLQI > 5	Post-baseline DLQI total score <=5 with baseline total score >5	Missing if baseline is missing or <=5 or observed value is missing
		DLQI total score in (0,1)	Post-baseline DLQI total score in (0,1)	N/A – partial assessments cannot be saved.
		4-point DLQI improve-ment in subgroup of patients with baseline DLQI total score ≥4	Change from baseline ≤ -4 and baseline ≥4	Missing if baseline is missing or <4 or observed value is missing.

Description and Derivation of Health Outcomes and Quality-of-Life Measures

Measure	Description	Variable	Derivation / Comment	Imputation Approach if with Missing Components
Work Productivity and Activity Impairment: Atopic Dermatitis (WPAI-AD)	The 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 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.
		Change from baseline in work impairment	Change from baseline: observed work impairment – baseline work impairment	Missing if baseline or observed value is missing.

Description and Derivation of Health Outcomes and Quality-of-Life Measures

Measure	Description	Variable	Derivation/Comment	Imputation Approach if with Missing Components
Work Productivity and Activity Impairment: Atopic Dermatitis (WPAI-AD)	The 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.	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.

Description and Derivation of Health Outcomes and Quality-of-Life Measures

Measure	Description	Variable	Derivation / Comment	Imputation Approach if with Missing Components
European Quality of Life–5 Dimensions–5 Levels (EQ-5D-5L)	The 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 labeled "best imaginable health state" and "worst imaginable health state."	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 state index)	Derive EQ-5D-5L UK Population-based index score according to the link by using the UK algorithm to produce a patient-level index score between -0.59 and 1.0 (continuous variable).	N/A – partial assessments cannot be saved on the eCOA tablet.

Description and Derivation of Health Outcomes and Quality-of-Life Measures

Measure	Description	Variable	Derivation / Comment	Imputation Approach if with Missing Components
European Quality of Life–5 Dimensions–5 Levels (EQ-5D-5L)	The 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 labeled "best imaginable health state" and "worst imaginable health state."	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.

Description and Derivation of Health Outcomes and Quality-of-Life Measures

Measure	Description	Variable	Derivation/Comment	Imputation Approach if with Missing Components
European Quality of Life–5 Dimensions–5 Levels (EQ-5D-5L) (continued)	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 [WWW]).	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.
		Change from baseline in sleep-wake and itch patterns	Change from baseline: observed score – baseline score	Missing if baseline or observed value is missing.

Abbreviations: AD = atopic dermatitis; eCOA = Electronic version of Clinical Outcome Assessment; EQ-5D = European Quality of Life–5 Dimensions; N/A = not applicable; QoL = quality of life; VAS = visual analog scale.

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

Measure	Variable	Analysis Method (Section 6.2.4)	Population (Section 6.2.2)	Comparison/Time Point	Analysis Type
Itch Numeric Rating Scale (NRS)	Itch NRS score Change from baseline in Itch NRS score	MMRM (primary censoring rule)	ITT	Bari 2-mg or Bari 1-mg vs PBO; Week 4, 16	Secondary Analysis
		ANCOVA using mLOCF (primary censoring rule)	ITT	Bari 2-mg or Bari 1-mg vs PBO; Week 4, 16	Sensitivity Analysis
	Percent change from baseline Itch score	MMRM (primary censoring rule)	ITT	Bari 2-mg or Bari 1-mg vs PBO; Week 1, 16	Secondary Analysis
		ANCOVA using mLOCF (primary censoring rule)	ITT	Bari 2-mg or Bari 1-mg vs PBO; Week 1, 16	Sensitivity Analysis
	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 (both censoring rules for ITT and primary censoring rule for PPS)	ITT	Bari 2-mg or Bari 1-mg vs PBO; Week 1, 2, 4, 16	Key Secondary Analysis
		Logistic regression using pMI and Tipping Point (primary censoring rule)	ITT	Bari 2-mg or Bari 1-mg vs PBO; Week 16	Sensitivity Analysis
		Cumulative Incidence Function of Time to reaching EASI 75% reduction (primary censoring rule)	ITT	Bari 2-mg, 1-mg vs PBO; Weeks 0 to 16	Exploratory analysis

Description of Health Outcomes and Quality-of-Life Measures Analyses

Measure	Variable	Analysis Method (Section 6.2.4)	Population (Section 6.2.2)	Comparison/Time Point	Analysis Type
Skin Pain Numeric Rating Scale (NRS)	Skin Pain NRS score Change from baseline in Skin Pain NRS score	MMRM NRI (both censoring rules for ITT and primary censoring rule)	ITT	Bari 2-mg or Bari 1-mg vs PBO; Week 16	Key Secondary Analysis /secondary censoring rule, sensitivity analysis
		ANCOVA using mLOCF (primary censoring rule)	ITT	Bari 2-mg or Bari 1-mg vs PBO; Week 16	Sensitivity Analysis
		ANCOVA using mBOCF (primary censoring rule)	ITT	Bari 2-mg or Bari 1-mg vs PBO; Week 16	Sensitivity Analysis
		pMI (primary censoring rule)	ITT	Bari 2-mg or Bari 1-mg vs PBO; Week 16	Sensitivity Analysis
	Skin pain 4-pt improvement in Skin pain NRS in subgroup of patients who had baseline Skin pain NRS ≥4	Logistic regression using NRI (primary censoring rule)	ITT	Bari 2-mg or Bari 1-mg vs PBO; Week 16	Exploratory Analysis

Description of Health Outcomes and Quality-of-Life Measures Analyses

Measure	Variable	Analysis Method (Section 6.2.4)	Population (Section 6.2.2)	Comparison/Time Point	Analysis Type
Atopic Dermatitis Sleep Scale (ADSS)	• ADSS item 2 scores Change from baseline in ADSS item 2 scores	MMRM (both censoring rules)	ITT	Bari 2-mg or Bari 1-mg vs PBO; Week 1, 16	Key Secondary Analysis
		ANCOVA using mLOCF (primary censoring rule)	ITT	Bari 2-mg or Bari 1-mg vs PBO; Week 1, 16	Sensitivity Analysis
		ANCOVA using mBOCF (primary censoring rule)	ITT	Bari 2-mg or Bari 1-mg vs PBO; Week 1, 16	Sensitivity Analysis for ADSS item 2
		pMI (primary censoring rule)	ITT	Bari 2-mg or Bari 1-mg vs PBO; Week 16	Sensitivity Analysis
	▪ 1.5 point improvement on Item 2 of ADSS	Logistic regression using NRI (primary censoring rule)	ITT	Bari 2-mg or Bari 1-mg vs PBO; Week 16	Exploratory Analysis
	• Change from baseline in ADSS item 1 scores • Change from baseline in ADSS item 3 scores	MMRM (primary censoring rule)	ITT	Bari 2-mg or Bari 1-mg vs PBO; Week 16	Exploratory Analysis
Patient-Oriented Eczema Measure (POEM)	• POEM score Change from baseline in POEM score	MMRM (primary censoring rule)	ITT	Bari 2-mg or Bari 1-mg vs PBO; Week 16	Secondary Analysis
		ANCOVA using mLOCF (primary censoring rule)	ITT	Bari 2-mg or Bari 1-mg vs PBO; Week 16	Sensitivity Analysis
	• POEM 4-pt improvement in subgroup of patients who had baseline POEM >= 4	Logistic regression using NRI (primary censoring rule)	ITT	Bari 2-mg or Bari 1-mg vs PBO; Week 16	Exploratory Analysis

Description of Health Outcomes and Quality-of-Life Measures Analyses

Measure	Variable	Analysis Method (Section 6.2.4)	Population (Section 6.2.2)	Comparison/Time Point	Analysis Type
Patient Global Impression of Severity–Atopic Dermatitis (PGI-S-AD)	• PGI-S-AD score Change from baseline in PGI-S-AD score	MMRM (censoring rule #1)	ITT	Bari 2-mg or Bari 1-mg vs PBO; Week 16	Secondary Analysis
		ANCOVA using mLOCF (primary censoring rule)	ITT	Bari 2-mg or Bari 1-mg vs PBO; Week 16	Sensitivity Analysis
Hospital Anxiety Depression Scale (HADS)	• HADS domain scores Change from baseline in HADS domain	MMRM (primary censoring rule)	ITT	Bari 2-mg or Bari 1-mg vs PBO; Week 16	Secondary Analysis
		ANCOVA using mLOCF (primary censoring rule)	ITT	Bari 2-mg or Bari 1-mg vs PBO; Week 16	Sensitivity Analysis
	• HADS total score	MMRM (primary censoring rule)	ITT	Bari 2-mg or Bari 1-mg vs PBO; Weeks 1 - 16	Exploratory Analysis
	• HADS Anxiety < 8 in subgroup of patients who had baseline HADS Anxiety >= 8	Logistic regression using NRI (primary censoring rule)	ITT	Bari 2-mg or Bari 1-mg vs PBO; Week 16	Exploratory Analysis
	• HADS Depression < 8 in subgroup of patients who had baseline HADS Depression >= 8	Logistic regression using NRI (primary censoring rule)	ITT	Bari 2-mg or Bari 1-mg vs PBO; Week 16	Exploratory Analysis
	• HADS Anxiety or Depression score <8 in subgroup of patients with baseline HADS Anxiety or Depression score ≥8	Logistic regression using NRI (primary censoring rule)	ITT	Bari 2-mg or Bari 1-mg vs PBO; Week 16	Exploratory Analysis

Description of Health Outcomes and Quality-of-Life Measures Analyses

Measure	Variable	Analysis Method (Section 6.2.4)	Population (Section 6.2.2)	Comparison/Time Point	Analysis Type
Dermatology Life Quality Index (DLQI)	• DLQI total score • Change from baseline in DLQI • Observed and change from baseline in domain scores -Symptoms and feelings -Daily activities -Leisure -Work and school -Personal relationships -Treatment	MMRM (primary censoring rule)	ITT	Bari 2-mg or Bari 1-mg vs PBO; Week 16	Secondary Analysis
		ANCOVA using mLOCF	ITT	Bari 2-mg or Bari 1-mg vs PBO; Week 16	Sensitivity Analysis
	• DLQI 4-pt improvement in subgroup of patients who had baseline DLQI >= 4	Logistic regression using NRI (primary censoring rule)	ITT	Bari 2-mg or Bari 1-mg vs PBO; Week 16	Exploratory Analysis
	• DLQI total score <=5 in subgroup of patients who had baseline DLQI > 5	Logistic regression using NRI (primary censoring rule)	ITT	Bari 2-mg or Bari 1-mg vs PBO; Week 16	Exploratory Analysis
	• DLQI (0,1)	Logistic regression using NRI (primary censoring rule)	ITT	Bari 2-mg or Bari 1-mg vs PBO; Week 16	Exploratory Analysis

Description of Health Outcomes and Quality-of-Life Measures Analyses

Measure	Variable	Analysis Method (Section 6.2.4)	Population (Section 6.2.2)	Comparison/Time Point	Analysis Type
Work Productivity and Activity Impairment: Atopic Dermatitis (WPAI-AD)	• Observed and Change from baseline in employment status	Descriptive statistics (observed) (Secondary Censoring Rule)	ITT	No comparison: Week 16	Secondary Analysis
	Observed and Change from baseline in: • absenteeism • presenteeism • overall work impairment • impairment in activities	MMRM (primary censoring rule)	ITT	Bari 2-mg or Bari 1-mg vs PBO; Week 16	Secondary Analysis
		ANCOVA using mLOCF(primary censoring rule)	ITT	Bari 2-mg or Bari 1-mg vs PBO; Week 16	Sensitivity Analysis
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 (primary censoring rule)	ITT	Bari 2-mg or Bari 1-mg vs PBO: Week 16	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 (primary censoring rule)	ITT	Bari 2-mg or Bari 1-mg vs PBO; Week 16	Secondary Analysis
		ANCOVA using mLOCF (primary censoring rule)	ITT	Bari 2-mg or Bari 1-mg vs PBO; Week 16	Sensitivity Analysis

Abbreviations: ANCOVA = analysis of covariance; Bari = baricitinib; EQ-5D = European Quality of Life–5 Dimensions; ITT = intent-to-treat; mBOCF = modified baseline observation carried forward; mLOCF = modified last observation carried forward; MMRM = mixed model repeated measures; NRI = non-responder imputation; PBO = placebo; pMI=placebo multiple imputation; -= per protocol set; VAS = visual analog scale.

6.13. Bioanalytical and Pharmacokinetic/Pharmacodynamic Methods

No pharmacokinetic analyses are planned for this study.

6.14. Safety Analyses

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

Safety analyses will include data from first dose of the study treatment, before and after rescue, unless otherwise stated, and patients will be analyzed according to the investigational product to which they were randomized at Week 0 (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 treatmentemergent 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 treatment-emergent adverse events (TEAEs) and serious adverse events (SAEs). Safety analyses will use the safety population defined in Section 6.2.2.

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 Supplemental Form, and safety in special groups and circumstances, including adverse events of special interest (AESIs) (see Section 6.14.5), 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 severity), 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. Listings will include all safety data.

For the 16-week interim lock, all safety data up to the Week 16 visit and at time of the interim lock will be included in the safety analysis. Safety data from patients who permanently discontinued study drug prior to time of the interim lock will be included in the interim lock safety analysis up to 30 days post-last dose of the study drug.

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).

6.14.1. 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 reported in weeks using the following:

- ≥28 days, ≥56 days, ≥84 days, ≥112 days, ≥196 days, ≥364 days, ≥532 days, and ≥728 days
- >0 to <28 days, ≥28 days to <56 days, ≥56 days to <84 days, ≥84 days to <112 days, ≥112 days to <196 days, ≥196 days to <364 days, ≥364 days to <532 days, ≥532 days to 728 days, and ≥728 days

The exposure ranges for the 16-week interim lock will be up to 112 days and including any exposure greater or equal to 112 days.

Overall exposure 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

6.14.2. 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 TEAE (ie, the event is treatment-emergent if the severity is coded moderate or severe post-baseline and not treatment-emergent if the severity is coded mild post-baseline). If an event occurring post-baseline 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 in order to derive treatment emergence. Should there be insufficient data for AE start date to make this comparison (eg, 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 experiencing 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 given gender.

For events of interest, incidence rate (IR) per 100 patient-years of observation (PYO) may be provided for the final analysis. See the compound-level safety standards for more details.

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 2-mg group;
- by MedDRA PT with events ordered by decreasing frequency in the baricitinib 2-mg group.

6.14.2.1. 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 2-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 2-mg group for the common TEAEs. For each patient and TEAE, the maximum severity for the MedDRA level being displayed is the maximum post-baseline severity observed from all associated LLTs mapping to that MedDRA PT.

6.14.2.2. Serious Adverse Events

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

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

Important medical events that may not be immediately life threatening or result in death or hospitalization but may jeopardize the patient or may require intervention to prevent one of the other outcomes listed in the definition above.

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 2-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, regardless of when they occurred during the study, will also be provided.

6.14.2.3. 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 an AE

by treatment using MedDRA PT nested within SOC. Events will be ordered by decreasing frequency in the baricitinib 2-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 1 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.

6.14.2.4. Criteria for Notable Patients

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

6.14.3. 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 (eg, 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. Key results pertaining to the 4 key hepatic laboratory assessments (alanine aminotransferase [ALT], aspartate aminotransferase [AST], total bilirubin [TBL], 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 6.14.5.1).

There is 1 special circumstance for laboratory values to be derived based on regularly scheduled, protocol-specified analytes. The low-density lipoprotein (LDL)/high-density lipoprotein (HDL) ratio will be derived as the ratio of LDL cholesterol to HDL cholesterol. There are no central laboratory 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 post-baseline measure will be displayed in box plots for patients who have both a baseline and at least 1 post-baseline visit. The last nonmissing 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 (eg, 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. Box plots will be provided for the period of Week 0 to Week 16.
- Treatment-emergent high/low analyses: 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. These analyses will be provided for the period of Week 0 to Week 16 and for Week 0 to data cut.

For laboratory analyte measurements collected qualitatively, a listing of abnormal findings will be provided. The listing will include, but not be limited to, patient identifier (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 post-baseline.

The listing of specific reference ranges used in analysis of laboratory data will be provided.

Note that additional analyses of certain laboratory analytes will be discussed in Section 6.14.5.1 for hepatic analytes, Section 6.14.5.2 for analytes related to hematological changes, Section 6.14.5.3 for analytes related to lipids, Section 6.14.5.4 for analytes related to renal function, and Section 6.14.5.5 for creatinine phosphokinase (CPK).

6.14.4. Vital Signs and Other Physical Findings

For the 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 (eg, 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, diastolic blood pressure, 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 6.14.3 will be used to analyze the vital signs and physical characteristics.

Table JAIW.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 29 April 2015 as recommended by the Lilly Cardiovascular Safety Advisory Committee.

Table JAIW.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

6.14.5. 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 AESIs.

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.

6.14.5.1. Abnormal Hepatic Tests

Analyses for abnormal hepatic tests will involve 4 laboratory analytes: ALT, AST, TBL, and ALP. In addition to the analyses described in Section 6.14.3, 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 ≥3×, 5×, and 10× the central laboratory upper limit of normal (ULN) during the treatment period will be summarized for all patients with a post-baseline value and for subsets based on various levels of baseline.
    - The analysis of 3× ULN will contain 4 subsets: patients whose nonmissing maximum baseline value is ≤1× ULN, patients whose maximum baseline is >1× ULN but <3× ULN, patients whose maximum baseline value is ≥3× ULN, and patients whose baseline values are missing.
    - The analysis of 5× ULN will contain 5 subsets: patients whose nonmissing maximum baseline value is ≤1× ULN, patients whose maximum baseline is >1× ULN but <3× ULN, patients whose maximum baseline is ≥3× ULN but <5× ULN, patients whose maximum baseline value is ≥5× ULN, and patients whose baseline values are missing.
    - The analysis of 10× ULN will contain 6 subsets: patients whose nonmissing maximum baseline value is ≤1× ULN, patients whose maximum baseline is >1× ULN but <3× ULN, patients whose maximum baseline is ≥3× ULN but <5× ULN, patients whose maximum baseline is ≥5× ULN but <10× ULN, patients whose maximum baseline value is ≥10× ULN, and patients whose baseline values are missing.
- The percentages of patients with an AST measurement ≥3×, 5×, and 10× the central laboratory ULN during the treatment period will be summarized for all patients with a post-baseline 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 TBL measurement ≥2× the central laboratory ULN during the treatment period will be summarized for all patients with a post-baseline value and subset into 4 subsets: patients whose nonmissing maximum baseline value is ≤1× ULN, patients whose maximum baseline is >1× ULN but <2× ULN, patients whose maximum baseline value is ≥2× ULN, and patients whose baseline values are missing.

- The percentages of patients with an ALP measurement ≥1.5× the central laboratory ULN during the treatment period will be summarized for all patients with a post-baseline value and subset into 4 subsets: patients whose nonmissing maximum baseline value is ≤1× ULN, patients whose maximum baseline is >1× ULN but <1.5× ULN, patients whose maximum baseline value is ≥1.5× 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 plot using maximum post-baseline ALT divided by ULN versus maximum post-baseline TBL 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 post-baseline ALT and TBL contributes 1 point to the plot. The measurements do not need to be taken at the same blood draw. Symbols will be used to indicate randomized treatment.

When criteria are met for hepatic evaluation and completion of the hepatic safety case reporting form (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, 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).

6.14.5.2. Hematologic Changes

Hematologic changes will be defined based on clinical laboratory assessments. Common Terminology Criteria for Adverse Events (CTCAE) AEs will be applied for laboratory tests potentially related to myelosuppressive events 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 [WWW], 2010 [WWW]).

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 post-baseline CTCAE grade from worst baseline grade
- increase to Grade 1 or above at worst post-baseline
- increase to Grade 2 or above at worst post-baseline
- increase to Grade 3 or above at worst post-baseline
- increase to Grade 4 at worst post-baseline

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 post-baseline results will be presented by treatment group for each treatment period within the following categories:

- Decreased: post-baseline category < baseline category
- Increased: post-baseline category > baseline category
- Same: post-baseline 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 post-baseline value >600 billion/L (Lengfelder et al. 1998). Planned and unplanned measurements will be included.

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

6.14.5.3. 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 6.14.3 and with TEAEs potentially related to hyperlipidemia.

Categorical analyses will be performed using National Cholesterol Education Program (NCEP) Adult Treatment Panel III guidelines (NCEP 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 post-baseline results will be presented by treatment group for each treatment period within the following categories:

- Decreased: post-baseline category more desirable than baseline category
- Increased: post-baseline category less desirable than baseline category
- Same: post-baseline 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'
- increase to categories 'High' or 'Very high'
- increase to 'Very high'

Treatment-emergent abnormal HDL cholesterol will be characterized as

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

The percentages of patients with treatment-emergent potential hyperlipidemia will be summarized by treatment group, ordered by decreasing frequency in the baricitinib 2-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].

6.14.5.4. Renal Function Effects

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

Common Terminology Criteria for Adverse Events 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. Refer to the compound-level safety standards for details.

6.14.5.5. Elevations in Creatinine Phosphokinase

Elevations in CPK will be addressed using CTCAE criteria 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. Refer to the compound-level safety standards for details.

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 compound-level safety standards.

6.14.5.6. 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.

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 and by the data cut will also be summarized in 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.

The IR and 95% CI will be calculated for the overall observation time for infections of special interest (serious infections, treatment-emergent herpes zoster, treatment-emergent tuberculosis, treatment-emergent potential opportunistic infections [POIs]) for the final analysis.

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

The TEAE infections will be further analyzed in terms of POI, herpes zoster, and herpes simplex. A summary of hepatitis B virus (HBV) deoxyribonucleic acid (DNA) monitoring results and association between infection and neutropenia/lymphopenia will also be provided in the context of infections.

Opportunistic Infection

To identify POIs, the following approach will be used to identify 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 and colleagues (2015) definitions for OI.

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 nonmultidermatomal involvement of the primary and/or adjacent dermatomes only
    - complicated – documented ocular (cornea or deeper structure; eg, iritis, keratitis, retinitis, etc.) or motor nerve involvement (eg, palsy; postherpetic neuralgia does not meet criteria for motor nerve involvement)

    - uncomplicated, localized or nonmultidermatomal cases that are not complicated
- multidermatomal involvement beyond primary and adjacent dermatomes (ie, >3 contiguous dermatomes) or involvement of 2 or more noncontiguous dermatomes
    - complicated-documented ocular (cornea or deeper structure; eg, iritis, keratitis, retinitis, etc.) or motor nerve involvement
    - uncomplicated-multidermatomal cases
- disseminated-systemic infection, visceral or widespread cutaneous (eg, ≥5 dermatomes or 3 to 4 dermatomes including at least 1 noncontiguous)
- recurrent: >1 infection occurring in an individual patient during the course of participation in the baricitinib clinical program.

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

A summary table of herpes zoster will be provided based on the above classification. 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. Of note, in the context of herpes zoster, antiviral medication treatment is defined as medication that 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 is not limited to, aciclovir, brivudine, cidofovir, famciclovir, foscarnet, ganciclovir, penciclovir, valaciclovir, valganciclovir, vidarabine (best presented by J05AB, J05AC, J05AE, and J05AH ATC 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 compound-level safety standards (both narrow and broad terms in the herpes simplex section). 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 will be provided. Skin infections may be defined based on MedDRA PT (see the compound-level safety standards).

HBV DNA

A listing of patients with detectable HBV DNA post-baseline will be provided.

Hepatitis B virus DNA status post-baseline (not detectable, detectable but not quantifiable [ie, < lower limit of detection (LLOD)], quantifiable [ie, ≥ 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 or (2) CTCAE Grade 3 or greater. Infection events with onset date ≤14 days before or after the Grade 2 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.

6.14.5.7. Major Cardiovascular Events and other Cardiovascular Events

Potential major 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:

- major cardiovascular events
  - cardiovascular death
  - myocardial infarction
  - stroke
- other cardiovascular events
  - transient ischemic attack
  - hospitalization for unstable angina
  - hospitalization for heart failure
  - serious arrhythmia
  - resuscitated sudden death
  - cardiogenic shock
  - coronary revascularization (such as coronary artery bypass surgery or percutaneous coronary intervention)
- noncardiovascular 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 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, noncardiovascular 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.

6.14.5.8. Venous and Pulmonary Artery Thromboembolic Events

Events identified as representative of venous thromboembolic event (VTE) disease will be further classified as deep vein thrombosis (DVT), pulmonary embolism (PE), or other peripheral venous thrombosis and will be analyzed. The following definitions apply:

- deep vein thrombosis: 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 noncompressible venous segment on ultrasound or confirmation by other imaging modality (eg, computed tomography [CT] scan, magnetic resonance imaging [MRI]).
- pulmonary embolism: 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 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 nonsuperficial 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 6.14.5.7 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 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.

6.14.5.9. Arterial Thromboembolic (ATE) Events

Refer to the compound-level safety standards.

6.14.5.10. Malignancies

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

All the cases identified by malignant tumors SMQ will be assessed thorough Medical (Global Patient Safety/Business Unit)/ medical team 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 Lilly Safety System report, disregarding the length of gap between the last treatment dose date and the event date. The NMSC flag will be confirmed by the medical team 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.

6.14.5.11. Allergic Reactions/Hypersensitivity

A search will be performed using the current MedDRA version 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. An algorithmic case 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 2 of the following categories from the SMQ:
    - Category B - (Upper Airway/Respiratory)
    - Category C - (Angioedema/Urticaria/Pruritus/Flush)
    - Category D - (Cardiovascular/Hypotension).

In the present studies, where study drug is administered daily, events will be considered as associated with an individual administration of study drug when the events are reported on the same study day.

Events that satisfy the queries will be listed, by temporal order within patient ID, and will include SOC, PT, SMQ event categorization including detail on the scope (narrow, algorithmic, or broad), reported AE term, AE onset and end dates, severity, seriousness, outcome, etc.

In addition, summary tables will be provided. Refer to the compound-level safety standards for details.

6.14.5.12. 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 by the medical safety team. Each case will be assessed to determine whether it is a GI perforation. A summary table based on medical review may be provided and treatment comparisons will be made using Fisher's exact test.

6.14.5.13. 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 (ie, 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 may be provided. Refer to the Compound safety level standards for details.

6.14.5.13.1. Self-Harm Supplemental Form and Self-Harm Follow-up Form

The Self-Harm Supplemental 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 given on the Self-Harm Follow-Up Form will be provided.

6.15. Subgroup Analyses

Subgroup analyses comparing each dose of baricitinib to placebo will be performed on the ITT population at Week 16, with data up to rescue 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)
  - 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. The subgroup analyses for categorical outcomes 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 100(1-alpha)% CI of the difference in percentages using the Newcombe-Wilson without continuity correction will be reported. The 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.

6.16. Protocol Violations

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 the interim lock period (prior to the primary endpoint [Week 16]) with the potential to affect efficacy analyses will result in exclusion from the PPS population.

Potential examples of deviations include patients who receive excluded concomitant therapy, significant noncompliance 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 IPDs 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 population will be provided by treatment group.

6.17. Interim Analyses and Data Monitoring

A DMC will oversee the conduct of this trial. 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, but not limited to, study discontinuation data, AEs including SAEs, clinical laboratory data, and vital sign data. 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 Data Monitoring Committee Charter for Phase 3 Studies of Baricitinib in Atopic Dermatitis, Alopecia Areata and Systemic Lupus Erythematoosus Programs and further details are given in the Interim Analysis Plan in Section 6.17.1.

Besides DMC members, a limited number of preidentified individuals may gain access to the limited unblinded data, as specified in the unblinding plan, prior to the interim or final database lock, to initiate the final population pharmacokinetic/pharmacodynamic model development processes or for preparation of regulatory documents. Interim locks will be conducted after all patients have completed Week 16 or discontinued from the study, and at various timepoints thereafter to support subsequent updates to regulatory agencies. Information that may unblind the study personnel will be managed according to the study unblinding plan.

6.17.1. Interim Analysis Plan

Analyses for the DMC will include listings and/or summaries of the following information:

- patient disposition, demographics, and baseline characteristics
- concomitant medications
- exposure
- adverse events, to include the following:
    - treatment-emergent adverse events
    - serious adverse events, including deaths
        - selected special safety topics
- clinical laboratory results
- vital signs
- Columbia Suicide Severity Rating Scale

Summaries will include TEAEs, SAEs, special topic AEs, and treatment-emergent high and low laboratory and vital signs in terms of counts, percentages, and IRs, where applicable. For

continuous analyses, box plots of laboratory analytes will be provided by time point and summaries will include descriptive statistics.

The DMC may request efficacy data if they feel there is value and to confirm a reasonable benefit/risk profile for ongoing patients in the studies. If efficacy data is requested, it will be mean change from baseline of EASI score. Further details are given in the DMC Charter.

6.18. Planned Exploratory Analyses

The planned exploratory analyses are described in Sections 6.11 and 6.12. 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 toolkit analyses, which may be produced, will also be documented in a supplemental SAP.

6.19. Annual Report Analyses

Annual report analyses, such as the Development Update Safety Report, will be documented in a separate analysis plan.

6.20. 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' AEs are summarized by treatment group and 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).
- Adverse event reporting is consistent with other document disclosures (eg, CSR, manuscripts).

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

7. Unblinding Plan

Refer to the blinding and unblinding plan document for details.

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.

BożekA, Reich, A. Assessment of Intra-and Inter-Rater Reliability of Three Methods for Measuring Atopic Dermatitis Severity: EASI, Objective SCORAD, and IGA. *Dermatology*. 2017;233(1):16-22.

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 [published correction appears in *Arch Dermatol*. 2005;141(3):381]. *Arch Dermatol.* 2004;140(12):1513-1519.

[CFR] 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 July 21, 2017.

[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. Published August 9, 2006. Accessed July 25, 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. Published May 28, 2009. Updated June 14, 2010. Accessed July 25, 2017.

[ETFAD] European Task Force on Atopic Dermatitis. Severity scoring of atopic dermatitis: the SCORAD index. Consensus Report of the European Task Force on Atopic Dermatitis. *Dermatology*. 1993;186(1):23-31.

EuroQol Group. EQ-5D-5L User Guide. Version 2.1. Available at: https://euroqol.org/wp-content/uploads/2016/09/EQ-5D-5L_UserGuide_2015.pdf. Accessed August 6, 2018.

Futamura M, Leshem YA, Thomas KS, Nankervis H, Williams HC, Simpson EL. A systematic review of Investigator Global Assessment (IGA) in atopic dermatitis (AD) trials: Many options, no standards. *J Am Acad Dermatol*. 2016;74(2):288-294.

Gillings D, Koch G. The application of the principle of intention-to-treat to the analysis of clinical trials. *Drug Inf Jl.* 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 July 21, 2017.

[ICH E9 R1] International Conference on Harmonisation: Estimands and Sensitivity Analysis in Clinical Trials. Available at: http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E9/E9R1EWG_Step2_Guideline_2017_0616.pdf. Accessed December 11, 2019.

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. Abstract P-59.

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.

Langley RG, Feldman SR, Nyirady J, van de Kerkhof P, Papavassilis C. The 5-point Investigator's Global Assessment (IGA) Scale: A modified tool for evaluating plaque psoriasis severity in clinical trials. *J Dermatolog Treat*. 2015;26(1):23-31.

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 Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III). 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. *Circulation*. 2002;106(25):3143-3421.

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 Qual Life Outcomes*. 2003;1:29.

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.

Leo Document ID = e17560a9-8d42-4c01-b179-004b09bca795

Approver: PPD

Approval Date & Time: 13-Dec-2019 19:22:21 GMT

Signature meaning: Approved