

Statistical Analysis Plan I4V-MC-JAHG

A Randomized, Double-Blind, Placebo-Controlled, Phase 2 Study to Evaluate the Safety and Efficacy of Baricitinib in Patients with Moderate-to-Severe Atopic Dermatitis

NCT02576938

Approved: 11 Jan 2017

### Electronic Signature Page

#### Workflow History

The following individuals have electronically signed this document.

Assigned To	Purpose	Date Signed	Status	Outcome
	Approve	11 Jan 2017 11:47:12 AM	Completed	Approved
	Author	11 Jan 2017 11:12:58 AM	Completed	Author
	Review	11 Jan 2017 11:24:07 AM	Completed	Reviewed
	Review	11 Jan 2017 11:35:19 AM	Completed	Reviewed
	Review	11 Jan 2017 11:37:44 AM	Completed	Reviewed

*Confidential*

**Statistical Analysis Plan  
I4V-MC-JAHG**

**A Randomized, Double-Blind, Placebo-Controlled, Phase 2  
Study to Evaluate the Safety and Efficacy of Baricitinib in  
Patients with Moderate-to-Severe Atopic Dermatitis**

Version 1.0

10 January, 2017

Prepared for

**Lilly | CHORUS**

**Eli Lilly and Company**

Lilly Corporate Center, Indianapolis IN 46285 U.S.A

Prepared by



55 Corporate Woods  
9300 West 110th Street, Suite 550  
Overland Park, Kansas 66210

## Table of Contents

Description	Page
<b>ABBREVIATIONS/DEFINITIONS .....</b>	<b>5</b>
<b>1 INTRODUCTION AND OBJECTIVES.....</b>	<b>7</b>
1.1 Introduction .....	7
1.2 Study Objectives.....	7
1.3 Determination of Sample Size.....	9
<b>2 General Statistical Methodology and Conventions .....</b>	<b>9</b>
2.1 Randomization Schedule and Unblinding Plan.....	9
2.2 Analysis Populations .....	10
2.3 Handling of Dropouts and Missing Data .....	10
2.4 Adjustment for Multiple Centers .....	11
2.5 Interim Analysis and Adjustment for Multiplicity .....	11
2.6 Coding of Concomitant Medications and Adverse Events.....	12
2.7 Definition of Study Time Points.....	12
2.8 Reporting Conventions .....	12
2.9 Changes to the Planned Analyses .....	13
2.9.1 Changes to the Planned Analyses in the Protocol .....	13
<b>3 Patient Accounting and Disposition.....</b>	<b>13</b>
3.1 Patient Accounting.....	13
3.2 Study Disposition.....	13
<b>4 Baseline Characteristics.....</b>	<b>13</b>
4.1 Demographics .....	14
4.2 Baseline Anthropometrics .....	14
4.3 Baseline Habits .....	14
4.4 Atopic Dermatitis History .....	14
<b>5 Concomitant Medications.....</b>	<b>15</b>
<b>6 Efficacy Analyses .....</b>	<b>15</b>
6.1 Primary Efficacy Analysis.....	15
6.1.1 Analyses in Support of Primary Efficacy.....	16
6.1.2 Sensitivity Analyses for the Primary Efficacy Variable .....	16
6.2 Secondary Efficacy Analyses.....	17

<b>6.2.1</b>	<b>Change and Percentage Change in EASI.....</b>	<b>17</b>
<b>6.2.2</b>	<b>SCORing Atopic Dermatitis (SCORAD) .....</b>	<b>17</b>
<b>6.2.3</b>	<b>Investigator’s Global Assessment .....</b>	<b>18</b>
<b>6.2.4</b>	<b>Dermatology Life Quality Index .....</b>	<b>18</b>
<b>6.2.5</b>	<b>Itch Numerical Rating Scale.....</b>	<b>19</b>
<b>6.3</b>	<b>Exploratory Efficacy Analyses.....</b>	<b>20</b>
<b>6.3.1</b>	<b>Patient-Oriented Eczema Measure.....</b>	<b>20</b>
<b>6.3.2</b>	<b>Analyses for Actigraphy Device data .....</b>	<b>21</b>
6.3.2.1	Analyses for Sleep/Wake Patterns.....	21
6.3.2.2	Analyses for Nocturnal Scratching.....	21
<b>6.3.3</b>	<b>Analyses for Quick Inventory of Depressive Symptomatology – Self Report 16 .....</b>	<b>21</b>
<b>6.3.4</b>	<b>Exploratory Subgroup Assessment of Response .....</b>	<b>22</b>
<b>7</b>	<b>Safety Analyses .....</b>	<b>22</b>
<b>7.1</b>	<b>Study Medication Exposure and Treatment Compliance .....</b>	<b>23</b>
<b>7.1.1</b>	<b>Extent of Exposure .....</b>	<b>23</b>
<b>7.1.2</b>	<b>Treatment Compliance .....</b>	<b>23</b>
<b>7.2</b>	<b>Adverse Events .....</b>	<b>24</b>
<b>7.2.1.</b>	<b>Treatment-Emergent Adverse Events .....</b>	<b>24</b>
<b>7.2.2.</b>	<b>Treatment Related TEAE.....</b>	<b>24</b>
<b>7.2.3</b>	<b>Serious Treatment-Emergent Adverse Events .....</b>	<b>24</b>
<b>7.2.4</b>	<b>TEAE Resulting in Death .....</b>	<b>25</b>
<b>7.2.5</b>	<b>TEAE Leading to Study Drug Discontinuation.....</b>	<b>25</b>
<b>7.2.6</b>	<b>TEAE by Maximal Severity .....</b>	<b>25</b>
<b>7.2.7</b>	<b>Adverse Events of Special Interest .....</b>	<b>25</b>
<b>7.3</b>	<b>Clinical Laboratory Results .....</b>	<b>25</b>
<b>7.3.1</b>	<b>Descriptive Statistics over Time.....</b>	<b>26</b>
<b>7.3.2</b>	<b>Shift tables and Categorical variables.....</b>	<b>26</b>
<b>7.4</b>	<b>Vital Signs and Weight .....</b>	<b>26</b>
<b>7.5</b>	<b>Biomarker Results.....</b>	<b>27</b>
<b>7.5.1</b>	<b>Skin biopsies.....</b>	<b>27</b>
<b>7.5.2</b>	<b>Blood biomarkers .....</b>	<b>27</b>
<b>7.5.3</b>	<b>Filaggrin genotyping .....</b>	<b>27</b>

**7.5.4 Correlation between Baseline Biomarkers .....27**

**7.5.5 Biomarkers as Measures of Disease Severity.....28**

**8 Pharmacokinetic Analyses..... 28**

**9 Health Outcome Psychometric Analyses..... 28**

**References..... 29**

## ABBREVIATIONS/DEFINITIONS

### Abbreviations

AE	adverse event
ALP	alkaline phosphatase
ALT	alanine aminotransferase
ANCOVA	analysis of covariance
ANOVA	analysis of variance
AST	alanine aminotransferase
BMI	body mass index
CMH	Cochran-Mantel-Haenszel
CSR	clinical study report
DLQI	Dermatologic Life Quality Index
eCRF	electronic case report form
EASI	Eczema Area and Severity Index
EASI-50	50% reduction in EASI score
EASI-75	75% reduction in EASI score
EASI-90	90% reduction in EASI score
GPORWE	(Lilly) Global Patient Outcomes and Real World Evidence
IFN-g	interferon-gamma
IGA	Investigator's Global Assessment
IGA [0]	post-baseline IGA Score of 0 (clear)
IGA [0,1]	post-baseline IGA Score of 0 (clear) or 1 (almost clear)
IRT	Interactive response technology
Itch NRS	Itch Numeric Rating Scale
ITT	intent to treat
LDH	lactate dehydrogenase
LOCF	last observed carried forward
LY	baricitinib (LY3009104)
MedDRA™	Medical Dictionary for Regulatory Activities
MMRM	mixed effects model repeated measures
NRI	non-responder imputation
PK	pharmacokinetic
POEM	Patient-Oriented Eczema Measure

PT	preferred term
QIDS-SR16	Quick Inventory Depressive Symptomatology – Self Report 16
QOL	quality of life
SAE	serious adverse event
SAP	Statistical analysis plan
SCORAD	SCORing Atopic Dermatitis
SDTM	Study Data Tabulation Model
SOC	system organ class
TBili	total bilirubin
TEAE	treatment emergent adverse event
ULN	upper limit of the normal range
VAS	visual analog scale
WASO	Wake after sleep onset
WHODD	World Health Organization Drug Dictionary



## 1 INTRODUCTION AND OBJECTIVES

### 1.1 Introduction

The purpose of this statistical analysis plan (SAP) is to describe the analysis variables and statistical procedures that will be used to analyze and report the results from a Phase 2 study evaluating the safety and efficacy of baricitinib, a JAK1/JAK2 selective inhibitor, in patients with moderate-to-severe atopic dermatitis. This SAP is based on the amended protocol, Protocol I4V-MC-JAHG(d), which was approved on 24 June 2016.

Changes to the protocol that impact the design, the data collected, or the statistical methods and that occur after the finalization of this SAP may require amendment of the approved SAP. Similarly, changes to the planned analysis variables and/or statistical methods described in the approved SAP may also require amendment of the SAP.

The formats for the tables, listings, and figures described in this SAP are provided in a companion document. Changes to the formats of these reports that are decided after the finalization of the SAP will not require an amendment. In addition, any additional supportive or exploratory analyses requested after SAP approval will not require amendment of the SAP. These additional analyses will be described in the clinical study report (CSR).

Please see the study protocol for details about the study design, procedures, and schedule of assessments and see the electronic case report form (eCRF) for details about variables collected and their possible values.

### 1.2 Study Objectives

[Table JAHG.1](#) shows the objectives and endpoints of the study that were defined in the protocol. Additional endpoints will be defined in this SAP.

**Table JAHG.1. Objectives and Endpoints**

Objectives	Endpoints
<p><b>Primary</b></p> <ul style="list-style-type: none"> <li>To compare the proportion of moderate-to-severe atopic dermatitis patients achieving a 50% or greater reduction in the Eczema Area and Severity Index (EASI-50) between each baricitinib dose group (2 mg and 4 mg) and placebo when treated daily for 16 weeks</li> </ul>	<ul style="list-style-type: none"> <li>EASI-50 at 16 weeks</li> </ul>
<p><b>Secondary</b></p> <ul style="list-style-type: none"> <li>To evaluate the absolute and percent change from baseline of the EASI with baricitinib compared to placebo</li> </ul>	<ul style="list-style-type: none"> <li>EASI at Weeks 1, 4, 8, 12, and 16</li> </ul>

<ul style="list-style-type: none"> <li>• To evaluate the mean change from baseline compared to placebo for the SCORing Atopic Dermatitis (SCORAD)</li> <li>• To evaluate the mean change from baseline compared to placebo for the Investigator’s Global Assessment (IGA)</li> <li>• To assess quality of life based on the Dermatologic Life Quality Index (DLQI)</li> <li>• To assess itch based on the Itch Numerical Rating Scale (NRS)</li> <li>• To characterize the pharmacokinetics of baricitinib in patients with moderate-to-severe atopic dermatitis</li> </ul> <p><b>Exploratory</b></p> <ul style="list-style-type: none"> <li>• To evaluate EASI response in patients based on their baseline total serum IgE</li> <li>• To evaluate total serum IgE over the course of treatment</li> </ul> <p><b><u>Exploratory (continued)</u></b></p> <ul style="list-style-type: none"> <li>• To evaluate changes in disease activity over the course of treatment</li> <li>• To evaluate changes in sleep quality over the course of treatment</li> <li>• To evaluate changes in nocturnal itch patterns over the course of treatment</li> <li>• To evaluate EASI response relative to filaggrin genotype</li> <li>• To assess tissue biomarkers from biopsies</li> <li>• To assess peripheral biomarkers mechanistically related to inhibition of JAK 1/2 mechanism</li> </ul>	<ul style="list-style-type: none"> <li>• SCORAD at Weeks 1, 4, 8, 12, and 16</li> <li>• IGA at Weeks 1, 4, 8, 12, and 16</li> <li>• DLQI at Weeks 1, 4, 8, 12, and 16</li> <li>• Itch Numerical Rating Scale at Weeks 1, 4, 8, 12, and 16</li> <li>• Plasma pharmacokinetic data</li> <li>• EASI at Weeks 1, 4, 8, 12 and 16; baseline total serum IgE</li> <li>• Total serum IgE at baseline, Weeks 4, and 16</li> <li>• Patient-Oriented Eczema Measure (POEM) at Weeks 1, 4, 8, 12, and 16</li> <li>• Change from baseline with data from a wearable actigraphy device</li> <li>• Change from baseline with data from a wearable actigraphy device</li> <li>• EASI-50 and filaggrin genotype data from peripheral baseline blood sample</li> <li>• K16, Ki-67, and IL-4 will be assessed from each tissue biopsy (collected at baseline (lesion and non-lesion), at Week 4 (lesion), and at Week 16 (lesion and non-lesion))</li> <li>• Peripheral biomarkers collected at baseline, Weeks 4 and 16</li> </ul>
--	---

<ul style="list-style-type: none"> <li>To perform subgroup assessments of response based on patient and disease characteristics</li> </ul>	<ul style="list-style-type: none"> <li>Ad hoc evaluation of potential prognostic and predictive disease characteristics</li> </ul>
--	--

### 1.3 Determination of Sample Size

Approximately 120 patients will be enrolled using a 4:3:3 randomization scheme for placebo, 2 mg baricitinib, or 4 mg baricitinib, respectively. Assuming a drop-out rate no greater than 20%, approximately 96 patients will complete the study.

This sample size provides 90% power to detect a 35% difference in EASI-50 between placebo and either the 4-mg baricitinib using a 2-sided chi-square test with  $\alpha = 0.05$ .

## 2 General Statistical Methodology and Conventions

CHORUS will designate a provider to generate the statistical analyses detailed in this SAP. All computations for statistical analyses will be performed using SAS<sup>®</sup> software, Version 9.3 or later. All SAS programs used in the production of statistical summary outputs will be validated with independent programming prior to finalization. In addition, all program outputs will be independently reviewed. The validation process will be used to confirm that all data manipulations and calculations were accurately done. Once validation is complete, a senior statistical reviewer should perform a final review of the documents to ensure the accuracy and consistency with this plan and consistency within tables. Upon completion of validation and quality review procedures, all documentation will be collected and filed by the project statistician or designee.

Before implementation of parametric methods of analysis, the distribution of analysis variables will be examined to determine if model assumptions are satisfied. Transformations or nonparametric methods of analysis may be used if warranted. However, in some cases, nonparametric analysis may be the initially proposed method due to the expected distribution of response. Whenever alternative methods of analysis are required, the description of the new method along with the rationale for its use will be documented in the CSR.

The eCRF data for all patients will be provided in Standard Data Tabulation Model (SDTM) datasets. Any additional data listing supplied as part of the CSR will be sorted by investigative site and patient identification number, and patients will be identified in the listings by the investigator number concatenated with the patient number.

### 2.1 Randomization Schedule and Unblinding Plan

This is a double-blind study in which the patients, investigator, and study site personnel will be blinded to treatment allocation. PAREXEL International will be responsible for building the interactive response technology (IRT) system that randomly assigns study treatments to patients.

To preserve the blinding of the study, only a minimum number of Lilly personnel will see the randomization table and treatment assignments before the study is complete. A limited group may have access to unblinded treatment information to ensure the safety of patients in the study. Another small team of Lilly scientists (project statistician, statistical analyst, pharmacokineticist,

research physician, medical director, and global patient outcome scientist) will have access to the unblinded treatment information just before the last patient has completed their Week 12 visit. Their review of the unblinded data will facilitate planning of future studies. They may communicate unblinded findings to a small number of other Lilly personnel to facilitate business decisions, but not to those who are involved in the conduct of the study who would be involved in direct site contact and assessment of blinded clinical outcomes. A list of all those individuals who become unblinded prior to study completion will be maintained by the unblinded statistician. This process will ensure that data integrity is maintained while patients are still receiving blinded therapy.

Once the last patient completes Week 16, thus blinded study treatment, and all data management data activities have been completed – data entered, coding completed, and all queries resolved – the Chorus Asset Manager and Statistician will approve PAREXEL to release the randomization schedule to the reporting team. Additional Lilly personnel as well as steering committee members may also be unblinded to commence reporting of results from the treatment period. Site personnel and those Lilly personnel who are involved in the conduct of the study who would be involved in direct site contact and assessment of blinded clinical outcomes and patients will remain blinded until the study completes.

Emergency unblinding of site personnel may be performed through the IRT system. This option may be used ONLY if the patient's well-being requires knowledge of the patient's treatment assignment. All calls resulting in an unblinding event are recorded and reported by the IRT system.

## 2.2 Analysis Populations

The analyses for efficacy will be based on the intention-to-treat (ITT) principle (Gillings and Koch 1991), which seeks to preserve the benefits of randomization and avoid the issue of selection bias. The **ITT population** is defined as all randomized patients, and analyses based on the ITT population will utilize the treatments assigned at randomization.

Safety will be analyzed using the **Safety Population**. The Safety population is defined as all randomized patients who received at least one dose of a randomly assigned study treatment. Safety analyses will utilize the treatment that the patient received.

Pharmacokinetic (PK) data will be analyzed using the **PK Population**. The PK population is defined as all patients who received at least one dose of baricitinib and provided at least one post-dose PK sample. Pharmacokinetic analyses will be performed using the treatment that the patient received.

## 2.3 Handling of Dropouts and Missing Data

Analysis of the primary response endpoint, EASI-50, will include all randomized patients and utilize a non-responder imputation (NRI) analysis. All patients who discontinue study treatment at any time prior to the time point of interest, or discontinue from the study for any reason, will be defined as non-responders for the NRI analysis. Randomized patients without at least 1 post-baseline observation will also be defined as non-responders for the NRI analysis. Secondary and exploratory efficacy responder analyses will also utilize NRI analyses and all randomized patients.

For the secondary and exploratory efficacy endpoints and clinical laboratory endpoints that are collected longitudinally, a last observation analysis will be performed by carrying forward the last post-baseline assessment. These last observation carried forward (LOCF) analyses guarantee the maximum number of randomized patients who were assessed post-baseline, will be included in the analyses. In addition, for many of the efficacy endpoints, mixed effect model repeated measures (MMRM) analyses will be performed to mitigate the impact of missing data, which will be assumed to be missing at random during the study.

If only a missing or partial date for AEs or concomitant medications is available and a complete date is required for calculations, the following algorithms will be applied:

- For the start date:
  - If year, month, and day are missing then use the minimum of the patient's first visit date or the consent date.
  - If either only month or month and day are missing then use January 1.
  - If only day is missing, impute the first day of the month.
- For the end date:
  - If year, month, and day are missing then use the patient's last visit date.
  - If either only month or month and day are missing then use December 31.
  - If only day is missing then use the last day of the month.
  - Do not expand the record past the patient's last visit.

The original missing or partial date, the imputed complete date, and the indicator variable that indicates which dates were imputed will be retained in the database.

Analyses of the safety endpoints, many of which are incidence based, will include all patients in the safety population, unless specifically stated otherwise.

## 2.4 Adjustment for Multiple Centers

The study randomization is stratified by country of the investigative site, and only investigative centers from Japan and the United States are included in the study. The number of patients to be randomized in each country is not fixed, so there may be low enrollment in one of the countries. The stratification variable “country” will not be included in the statistical model by default. If there are at least 15 patients enrolled in each country, analyses that include a fixed effect for country may be performed as exploratory analyses.

## 2.5 Interim Analysis and Adjustment for Multiplicity

There were no planned formal interim analyses for this study. The interim review of the ongoing study data, that is described in [Section 2.1](#), does not include any alpha-spending nor will any modifications of the current study be contemplated.

The primary efficacy analyses include separate comparisons of each baricitinib arm to the placebo arm. To control the type 1 error rate for the primary efficacy analyses, these separate comparisons will be performed in a stepwise fashion. The 4-mg baricitinib arm will be compared with the placebo arm; if this test is statistically significant at level  $\alpha = 5\%$ , then the 2-mg baricitinib arm will be compared with the placebo arm; again using  $\alpha = 5\%$ .

Because this is a phase 2 study, no other adjustments for multiplicity will be made for statistical tests.

## 2.6 Coding of Concomitant Medications and Adverse Events

Adverse events (AEs) will be coded using the Medical Dictionary for Regulatory Activities (MedDRA™), and concomitant medications will be coded using the World Health Organization Drug Dictionary. The most recent version of these dictionaries at the time of database lock will be used for reporting. The version of the dictionary used for reporting will be provided in the CSR.

## 2.7 Definition of Study Time Points

Study baseline for a patient will extend from the date of informed consent until the date of first dose of randomly assigned study medication. For efficacy, safety and background characteristics, the baseline value of a parameter will be the last measurement of that parameter during the baseline period.

The post-baseline period will start on the date of first dose of randomly assigned study medication. Study day will be defined as

$$\text{Study Day} = \text{Date} - \text{Date of 1}^{\text{st}} \text{ dose} + 1 * \text{if} (\text{Date} \geq \text{Date of 1}^{\text{st}} \text{ dose}).$$

The last part of the equation indicates that 1 is added if the date of collection/observation is greater than or equal to the date of first dose of randomly assigned study medication. This forces all post-baseline events to have a positive value and all dates prior to first dose to have a value less than 0.

## 2.8 Reporting Conventions

This section details the general conventions to be used for the statistical analyses. Departures from these general conventions will be provided in the specific detailed sections of this analysis plan. The following conventions will be applied to all data presentations and analyses:

- Data will be summarized for the following groups:
  - Overall, Placebo, 2mg LY, 4mg LY will be used for patient accounting and disposition
  - Placebo, 2mg LY, and 4mg LY for all presentations except patient accounting and disposition
- Continuous variables will generally be summarized by the number of patients, mean, standard deviation, median, minimum, and maximum. Exceptions to these conventions will be specifically noted. Categorical variables will be summarized by the number and percentage of patients within each category.
- All mean and median values will be formatted to one more decimal place than the measured value.
- Standard deviation values will be formatted to two more decimal places than the measured value.
- Minimum and maximum values will be presented with the same number of decimal places as the measured value.
- The number and percent of responses will be presented in the form XX (XX %), where the percentage is in parentheses. Percentages will be rounded to the nearest percent. In the case of a frequency of zero, the frequency and percentage will be presented as 0 rather than 0 (0%).

- All summary tables will include the analysis population sample size (i.e., number of patients) in each treatment group.
- Date variables will be formatted as ddMMYY for presentation.

## **2.9 Changes to the Planned Analyses**

### ***2.9.1 Changes to the Planned Analyses in the Protocol***

- There was a typo for the analyses of continuous secondary efficacy endpoints. The level for the confidence intervals should have been 95%, rather than 90%. The correction has been implemented in this SAP.

## **3 Patient Accounting and Disposition**

### **3.1 Patient Accounting**

For each study center the number of patients in the ITT, Safety and PK populations will be provided overall and by randomized assignment. The date of the first patient visit and the last patient visit for each study center will also be provided. These dates and patient totals will also be summarized for the whole study and by country.

A list of protocol violations that could potentially impact the analysis of the study will be determined during the conduct of the study by study team members who are blinded to assigned study treatment. The number and percentage of patients with each violation will be tabulated overall and by randomized treatment assignment.

### **3.2 Study Disposition**

The disposition of all randomized patients will be presented. The number of patients randomized and the number of patients in the Safety Population will be presented overall and by treatment group. In addition, the reason for study discontinuation will be tabulated by treatment group using the list of reasons provided in the eCRF. This report will be generated for both the overall population and for patients enrolled in Japan.

## **4 Baseline Characteristics**

Demographic data, baseline characteristics, and atopic dermatitis history will be summarized using descriptive statistics by randomized treatment assignment. These summaries will be based on the ITT population and will be generated using both the overall population and for patients enrolled in Japan. Randomized patients who are missing measurements of the baseline variable being analyzed will not be included in the summary for that variable.

Unless otherwise specified, F-tests from the analysis of variance (ANOVA) model with a fixed effect for randomized treatment group will be conducted for continuous variables. These tests will be based on the Type 3 sum of squares. Categorical variables will be analyzed using the general association chi-square tests. If the number of patients in a category is less than 5 for either randomized group, Freeman-Halton exact tests will be used instead of the chi-square test.



## 4.1 Demographics

Demographic variables collected prior to randomization include date of birth, sex, race, and ethnicity. Age at study entry, age group (Age < 65 years, Age ≥ 65 years), sex, and race will be summarized and reported. Age at study entry will be based on the age of the patient on the date the informed consent is signed. Age at study entry will be analyzed as a continuous variable while age group is analyzed as a categorical variable. Sex will also be analyzed as a categorical variable. Patient race will be captured using multiple racial categories (i.e., American Indian or Alaskan Native, Asian, Black or African American, Native Hawaiian or other Pacific Islander, and White), and Race will be reported using a chi-square test and these categories along with a 'Multiple' category for those patients who identify more than one race category. The response for ethnicity could be Hispanic or Latino, or Non-Hispanic or Latino. Ethnicity will be analyzed as a categorical variable.

## 4.2 Baseline Anthropometrics

Patient height and weight will also be collected prior to randomization. Both variables will be reported in metric units (height in cm and weight in kg) and will be analyzed as continuous variables along with Body Mass Index (BMI; in kg/m<sup>2</sup>). Body mass index will also be analyzed categorically using the categories obese (BMI ≥ 30 kg/m<sup>2</sup>) and non-obese (BMI < 30 kg/m<sup>2</sup>).

## 4.3 Baseline Habits

The alcohol, tobacco, and caffeine habits for each patient will be collected during the baseline period. The number and percentage of patients who consume alcohol, tobacco and caffeine will be tabulated for each treatment group using the ITT population. In addition, among those patients who currently smoke descriptive statistics of the number of cigarettes per day will be provided by treatment group. No statistical inference will be performed for the habits variables, and this report will only be produced for the overall population.

## 4.4 Atopic Dermatitis History

The date of atopic dermatitis diagnosis along with the types of prior therapies received will be collected during the screening period. The types of prior therapy will be collected using the following categories: hydration plus topical steroids and/or antibiotics (Category 1), systemic steroids and/or phototherapy (Category 2), cyclosporine and/or other immunomodulators (Category 3). Patients could have received treatments from more than one category.

Time since diagnosis of atopic dermatitis, in months, will be descriptively summarized and analyzed as a continuous variable. Since patients could have received treatments from multiple categories, each category will be analyzed separately as a categorical variable. In addition, the number and percent of patients who reported alopecia areata in their medical history will be compared between treatments using the general association Mantel-Haenszel or Freeman-Halton statistic.



## 5 Concomitant Medications

Concomitant medications will be defined as medications, other than triamcinolone 0.1% cream, taken on or after the date of first dose of randomly assigned study treatment. This includes all medications initially taken prior to the date of first dose of randomly assigned study medication but with a stop date that is either missing or after the date of first dose of randomly assigned study medication. Those medications where the stop date is documented as prior to the date of first dose of randomly assigned study medication will be classified as prior medications. The prior medications will not be included in any summary reports.

The reporting of concomitant medications will be by WHODD preferred drug name (often the generic drug name). The report of all concomitant medications taken during the study will be prepared by treatment using the safety population, and the number and percentage of patients who received each concomitant medication will be presented.

## 6 Efficacy Analyses

Efficacy analyses for this study will be based on the findings from clinical tools (EASI, SCORAD, and IGA) and patient reported outcomes (DLQI, Itch NRS, and POEM) that assess the extent and severity of atopic dermatitis, the symptoms related to atopic dermatitis, or the impact on the patient's quality of life (QOL). In addition, data from wearable actigraphy devices will be analyzed to assess sleep/wake patterns and changes in nocturnal itch patterns. All of these assessments will be collected longitudinally during the study.

All efficacy analyses will be performed using the ITT population. Reports will be generated for the overall population, and many of the reports will also be produced using only patients enrolled in Japan. The method of analysis or statistical technique employed to address missing data will be described separately for each variable. Presentations of efficacy will include the number of patients with data at each time point plus related statistics derived from the analysis.

Assessments performed at Visit 2 will be considered the baseline measurements. If the Visit 2 assessment for a clinical tool or patient reported outcome is missing, then the Visit 1 assessment, if it exists, will be considered the baseline measurement. If a patient has no assessment of a clinical tool or patient reported outcome prior to the initiation of study treatment, then the patient will be excluded from the efficacy analysis of that variable at the given time point.

### 6.1 Primary Efficacy Analysis

The primary efficacy analysis will be based on the Eczema Area and Severity Index (EASI), which is a valid and internally consistent clinical observation tool that has adequate intra-observer reliability, intermediate inter-observer reliability, and adequate responsiveness.

The EASI assesses objective physician estimates of two dimensions of atopic dermatitis – disease extent and clinical signs – 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 four clinical signs (erythema, induration/papulation, excoriation, and lichenification) each on a scale of 0 to 3 (0 = none, absent; 1 = mild; 2 = moderate; 3 = severe) at four body sites (head and neck, trunk, upper limbs, and lower limbs). Half scores are allowed. Each body site will have a score that ranges from 0 to 72, and the final EASI score will be obtained by weight-

averaging these four scores (using multipliers 0.2 for head and neck and upper limbs and 0.3 for trunk and lower limbs). Hence, the final EASI score will range from 0 to 72 for each time point.

The EASI will be assessed at baseline and at Weeks 1, 4, 8, 12, and 16 post-randomization, plus at the early termination visit, if applicable. At each of the post-baseline visits the EASI will be computed, and the EASI score will be compared with the EASI score obtained at baseline to obtain the proportion of reduction.

The primary efficacy measure will be the proportion of moderate-to-severe atopic dermatitis patients achieving a 50% reduction in the Eczema Area and Severity Index, or EASI-50, after 16 weeks of treatment. The primary analysis will be performed using a Non-Responder Imputation (NRI) analysis. All patients with a less than 50% reduction in the EASI at Week 16, plus all patients who either discontinue study treatment or discontinue from the study for any reason at any time prior to Week 16 will be defined as non-responders for the NRI analysis. Randomized patients without at least 1 post-baseline observation of EASI will also be defined as non-responders for the NRI analysis.

The proportion of patients achieving EASI-50 in the placebo arm will be compared to the proportion of patients in each of the baricitinib arms (2 mg and 4 mg) using a two-sided chi-square test with  $\alpha=0.05$ . To control for multiplicity testing will be performed in a stepwise fashion. First, the 4-mg baricitinib arm will be compared with the placebo arm; if this test is statistically significant, then the 2-mg baricitinib arm will be compared with the placebo arm.

The number and percent of patients, based on an NRI analysis, with a 50% reduction in EASI will be presented for each treatment arm at each time point including the patient's last assessment, which should be either Week 16 or the early termination visit. The p-values for the pairwise chi-square tests from the NRI analyses will also be tabulated. In addition, a bar chart that shows the percentage of patients with a 50% reduction in EASI for each treatment arm at each time point including the patient's last assessment will be presented. The figure displays will be generated for the overall population and by country – United States and Japan.

### ***6.1.1 Analyses in Support of Primary Efficacy***

Non-responder imputation analyses that assess the EASI-75 and EASI-90, i.e., proportion of moderate-to-severe atopic dermatitis patients achieving a 75% (or 90%) reduction in the EASI, will also be conducted. P-values from Fisher's exact test may be used due to the potential for low response rates.

Observed cases analyses of EASI-50, EASI-75 and EASI-90 will also be conducted using Chi-square of Fisher's Exact Test.

### ***6.1.2 Sensitivity Analyses for the Primary Efficacy Variable***

The NRI analysis of EASI-50, as well as EASI-75 and EASI-90, will be repeated using a logistic regression model with fixed effects for treatment and country. Odds ratio estimates and the Wald 95% confidence intervals will be provided for the pairwise comparisons of baricitinib to placebo.

## 6.2 Secondary Efficacy Analyses

### 6.2.1 *Change and Percentage Change in EASI*

Change and percentage change from baseline for EASI will be derived for each patient at each post-baseline visit (Weeks 1, 4, 8, 12, and 16). Descriptive statistics for the value, change and percentage change will be tabulated for each time point.

Missing assessments will assumed to be missing at random and separate mixed effects model repeated measures (MMRM) analyses will be performed to compare the baricitinib treatment arms and placebo for the actual and percent change from baseline. For both change and percent change from baseline the MMRM model will contain fixed effects for treatment, visit, and country and the treatment-by-visit interaction plus the baseline value and baseline-by-visit term as covariates. Patient will be nested within treatment, and the covariance structure among the repeated measurements for a patient will be assumed to be common across patients and modeled using an unstructured 5x5 matrix. Satterthwaite's approximation will be used to estimate denominator degrees of freedom. The significance of differences in least-square means will be based on Type III tests using the observed population margins. The analyses will be performed using the SAS procedure PROC MIXED.

The tabular output for these MMRM analyses will present the least squares means, standard errors and 95% confidence intervals and p-values for each main effect marginal. Pairwise contrasts of the least squares means between the randomized treatment arms will also be provided at each time point using LSMESTIMATE statements.

For both endpoints plots of the least squares mean  $\pm$  standard error estimates for changes in EASI over time will be provided by randomized treatment. These plots will be provided for the overall population and for each country.

Finally, a last observation analysis will be performed for both change and percent change using an analysis of covariance (ANCOVA) model with fixed effects for treatment and country plus the baseline value as a covariate. Tabular output for these ANCOVA analyses will present the least squares means, standard errors and 90% confidence intervals and p-values for each main effect marginal. Pairwise contrasts of the least squares means will also be provided for randomized treatment arms at each time point.

### 6.2.2 *SCORing Atopic Dermatitis (SCORAD)*

To determine disease severity the SCORing Atopic Dermatitis (SCORAD) index uses the rule of nines to assess disease extent (head and neck 9%; upper limbs 9% each; lower limbs 18% each; anterior trunk 18%; back 18%; genitals 1%; and hands 2%) and evaluates six clinical characteristics – erythema, edema/papulation, oozing/crusts, excoriation, lichenification and dryness – each on a scale of 0 to 3 (0 = none, absent; 1 = mild; 2 = moderate; 3 = severe). SCORAD also assesses subjective symptoms of pruritus and sleep loss with visual analogue scales (VAS) where 0 is no itch (or no sleeplessness) and 10 is the worst imaginable itch (or sleeplessness). These three aspects: extent of disease (A: range 0-102), disease severity (B; range 0-18) and subjective symptoms (C: range 0-20) will be combined using,  $A/5 + 7*B/2 + C$ , to give a maximum possible score of 103.4.

The SCORAD will be assessed at baseline and at Weeks 1, 4, 8, 12, and 16 post-randomization, plus at the early termination visit, if applicable. At each of the post-baseline visits the SCORAD will be computed, and the SCORAD score will be compared with the SCORAD score obtained at baseline to obtain change and percent change from baseline. Descriptive statistics for the SCORAD value and the subjective assessments of pruritus and sleep loss and the respective changes from baseline will be tabulated for each time point using the overall and by-country populations.

The MMRM and ANCOVA analyses described in [Section 6.2.1](#) will be performed to compare the baricitinib treatment arms and placebo for change and percent change from baseline in the SCORAD score, and change from baseline in the pruritus and insomnia assessments. The tabular outputs for the MMRM and ANCOVA analyses described will be provided.

### ***6.2.3 Investigator's Global Assessment***

The Investigator's Global Assessment (IGA) uses the clinical characteristics – erythema, infiltration, papulation, oozing and crusting – as guidelines for the overall severity assessment. The IGA consists of a 6-point severity scale from clear to very severe disease (0 = clear, 1 = almost clear, 2 = mild disease, 3 = moderate disease, 4 = severe disease and 5 = very severe disease).

The IGA will be assessed at baseline and at Weeks 1, 4, 8, 12, and 16 post-randomization, plus at the early termination visit, if applicable. At each of the post-baseline visits the IGA will be computed, and the IGA score will be compared with the IGA score obtained at baseline to obtain change from baseline. Descriptive statistics for the value and change from baseline will be tabulated for each time point using the overall and by-country populations.

An IGA [0,1] responder is defined as a patient with a post-baseline score of 0 (clear) or 1 (almost clear), and an IGA [0,1] responder with a 2-point drop is a responder whose change from baseline is at least 2 points. An IGA [0] responder is defined as a patient with a post-baseline score of 0 (clear). At each post-baseline time point and the last observation the number and percent of IGA [0,1] responders both overall and those with at least a 2-point drop from baseline will be compared between each baricitinib arm and placebo using a two-sided Fisher's Exact test at level  $\alpha = 0.05$ . A similar analysis will be performed for IGA [0].

In addition, a bar chart that shows the percentage of IGA [0,1] responders with a 2-point drop in each treatment arm and at each time point including the patient's last assessment will be presented. The figure will be generated for the overall population and by country.

Analyses will also be performed for change from baseline at each post-baseline time point and for the last observation. Comparisons between each baricitinib arm with placebo will be based on chi-square statistics using the row mean scores option. The number and percentage of patients having change from baseline scores (range: -5 to 5) will be tabulated for each treatment arm and each time point. The associated p-values at each time point will also be presented for each pairwise comparison.

### ***6.2.4 Dermatology Life Quality Index***

The Dermatology Life Quality Index (DLQI) is a simple, patient-administered, 10-question, 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 “not at all,” “a little,” “a lot,” and “very much,” with corresponding scores of 0, 1, 2, and 3 respectively and unanswered or “not relevant” responses scored as 0. For question 7 the response category ‘yes’ is scored 3, while a ‘no’ response is scored as 2, 1, or 0 based on the follow-up question response “a lot”, “a little”, or “not at all” respectively. The DLQI total score will be obtained by summing the scores from the 10 questions with a minimum score of 0 and a maximum score of 30. If any one of the 10 questions is unanswered, that question is scored as 0. However, if 2 or more questions are not answered then the DLQI score will not be calculated.

The DLQI will be assessed at baseline and at Weeks 1, 4, 8, 12, and 16 post-randomization, plus at the early termination visit, if applicable. At each of the post-baseline visits the DLQI total score will be computed along with the change from baseline. Descriptive statistics for the value, change from baseline will be tabulated for each time point.

The MMRM and ANCOVA analyses described in [Section 6.2.1](#) will be performed to compare the baricitinib treatment arms and placebo for change from baseline in the DLQI total score.

A DLQI responder is defined as a patient with a 4-point or greater decrease from baseline in the DLQI total score. At each post-baseline time point and the last observation the number and percent of DLQI responders will be compared between each baricitinib arm and placebo using a two-sided Fisher’s Exact test at level  $\alpha = 0.05$ . This analysis will be performed using only patients with a baseline DLQI total score greater than or equal to 4 points.

Individual domain DLQI scores may be analyzed in post-hoc analyses.

### **6.2.5 Itch Numerical Rating Scale**

The Itch Numerical 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.” The patient indicates the overall severity of itching that best describes the worst level of itching in the past 24 hours.

The Itch NRS will be assessed at baseline and at Weeks 1, 4, 8, 12, and 16 post-randomization, plus at the early termination visit, if applicable. At each of the post-baseline visits the Itch NRS score will be computed along with the change and percent change from baseline. Descriptive statistics for the value and changes from baseline will be tabulated for each time point.

The number and percent of patients with an Itch NRS score of 0 “no itch” will be compared between the baricitinib arms and the placebo arm at each time point including baseline using an NRI analysis. Patients who either discontinue study treatment or discontinue from the study for any reason will be defined as non-responders for the NRI analysis for all assessments that were not collected. Randomized patients without at least 1 post-baseline observation of Itch NRS will also be defined as non-responders for the NRI analysis. The proportion of patients with no itch in the placebo arm will be compared to the proportion of patients in each of the baricitinib arms (2 mg and 4 mg) using two-sided Fisher’s exact tests with  $\alpha=0.05$ . The number and percentage of patients with no itch will be presented for each treatment arm at each time point including the patient’s last assessment. The p-values for the pairwise Fisher’s Exact tests will also be tabulated.

In addition, a bar chart that shows the percent of patients with no itch for each treatment arm at each time point including the patient's last assessment will be presented. The figure display will be generated for the overall population and by country.

The MMRM and ANCOVA analyses described in [Section 6.2.1](#) will be performed to compare the baricitinib treatment arms and placebo for change from baseline in the Itch NRS scores. The tabular outputs for the MMRM and ANCOVA analyses described will be provided.

## 6.3 Exploratory Efficacy Analyses

### 6.3.1 Patient-Oriented Eczema Measure

The Patient Oriented Eczema Measure (POEM) is seven-item patient-administered tool that focuses on the atopic dermatitis experienced by the patient. This tool has a one week recall period. Each question is scored on a five point scale (0 = no days; 1 = 1-2 days; 2 = 3-4 days; 3 = 5-6 days; 4 = every day), with a total score ranging from 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.

The POEM score will be calculated at baseline and at Weeks 1, 4, 8, 12, and 16 post-randomization, plus at the early termination visit, if applicable. At each of the post-baseline visits the POEM score will be compared with the POEM score obtained at baseline to obtain change and percent change from baseline. Descriptive statistics for the value, change and percent change from baseline will be tabulated for each time point.

The MMRM and ANCOVA analyses described in [Section 6.2.1](#) will be performed to compare the baricitinib treatment arms and placebo for change from baseline in the POEM score. The tabular outputs for the MMRM and ANCOVA analyses described therein will be provided.

The POEM score can be banded as follows:

- 0 to 2 = Clear or almost clear
- 3 to 7 = Mild eczema
- 8 to 16 = Moderate eczema
- 17 to 24 = Severe eczema
- 25 to 28 = Very severe eczema

At each time point, baseline and post-baseline, the banded POEM score will be compared between each baricitinib arm and placebo using chi-square statistics with the row mean scores option. The number and percentage of patients in each band on the scale will be tabulated for each treatment arm and each time point. The associated p-values at each time point will be presented for each pairwise comparison. A bar chart that shows the distribution of patients among the bands for each treatment arm at each time point including the patient's last assessment will be presented.

Exploratory analyses may examine individual POEM items.



### **6.3.2 Analyses for Actigraphy Device data**

Hypoallergenic wrist watch-like devices will be worn by the patients throughout the study. A total of three devices will be worn – one device captures activity and sleep while two high-resolution devices, one worn on each arm, determine nocturnal scratching. Data from the devices will be used to summarize sleep/wake patterns and quantify nocturnal scratching events.

Data from the actigraphy devices will be downloaded at Weeks 1, 4, 8, 12, and 16 post-randomization, plus at the early termination visit, if applicable. Reproducibility of the data from day-to-day is fairly high, so results from the seven days prior to each download will be averaged to derive the response variables. Data from baseline will be available as part of the Week 1 download. At each of the post-baseline visits change from baseline will be computed for each of the response variables.

Additional exploratory analyses may be done, for example evaluating area under the curve to compare treatment groups.

#### **6.3.2.1 Analyses for Sleep/Wake Patterns**

At each time point the following parameters will be analyzed: average number of sleep disturbances per night-hour, duration of sleep, time of initial disturbance after sleep onset (WASO), onset latency, wake time, percent of time awake, number of wake bouts, fragmentation and sleep efficiency. The MMRM analysis described in [Section 6.2.1](#) will be performed to compare the baricitinib treatment arms and placebo for change from baseline for each of the variables. The tabular outputs for the MMRM analyses will be provided.

#### **6.3.2.2 Analyses for Nocturnal Scratching**

Three variables will be computed to assess nocturnal scratching – the average number of scratching events per night, the average time spent actively scratching during the night, and the average number of scratching events per hour. The MMRM analysis described in [Section 6.2.1](#) will be performed to compare the baricitinib treatment arms and placebo for change from baseline for each of the nocturnal itching variables. The tabular outputs for the MMRM analyses will be provided.

### **6.3.3 Analyses for Quick Inventory of Depressive Symptomatology – Self Report 16**

The Quick Inventory Depressive Symptomatology - Self Report 16 (QIDS-SR16) is a self-administered 16-item instrument intended to assess the existence and severity of symptoms of depression. The QIDS-SR16 scale is used to assess the potential impact of treatment on new onset or changes in depression, thoughts of death, and /or suicidal ideation severity. A patient is asked to consider each statement as it relates to the way they have felt for the past 7 days. There is a 4-point scale for each item ranging from 0 to 3. The 16 items corresponding to 9 depression domains are summed to give a single score ranging from 0 to 27, with higher scores denoting greater symptom severity. The domains assessed by the instrument include: (1) sad mood, (2) concentration, (3) self-criticism, (4) suicidal ideation, (5) interest, (6) energy/fatigue, (7) sleep disturbance (initial, middle, and late insomnia or hypersomnia), (8) decrease/increase in appetite/weight, and (9) psychomotor agitation/retardation. The score is derived by adding the

following nine numbers – results from questions 5 (depressed mood), 10 (concentration), 11 (worthlessness/guilt), 12 (suicidal ideation), 13 (decreased interest), 14 (decreased energy) plus the highest scores from questions 1-4 (sleep), 6-9 (weight/appetite changes), and 15-16 (psychomotor changes), respectively. If any single question is left unanswered, then that question is scored as 0.

The QIDS-SR16 will be assessed at baseline and at Weeks 1, 4, 8, 12, and 16 post-randomization, plus at the early termination visit, if applicable. At each of the post-baseline visits the QIDS-SR16 domain scores and total score will be calculated, and the scores will be compared with the respective baseline score to obtain change and percent from baseline. Descriptive statistics for the value and changes from baseline will be tabulated for each domain and total score and at each time point.

The QIDS-SR16 total score will be categorized as follows:

- 0 to 5 = No depression
- 6 to 10 = Mild depression
- 11 to 15 = Moderate depression
- 16 to 20 = Severe depression
- 21 to 27 = Very severe depression

At each time point, baseline and post-baseline, the level of depression will be compared between each baricitinib arm and placebo using chi-square statistics with the row mean scores option. The number and percentage of patients in each depression level will be tabulated for each treatment arm and each time point. The associated p-values at each time point will be presented for each pairwise comparison. A bar chart that shows the distribution of patients among the bands for each treatment arm at each time point including the patient's last assessment will be presented. The table and figure displays will be generated for the overall population and for patients enrolled in Japan.

### ***6.3.4 Exploratory Subgroup Assessment of Response***

Subgroup assessments of response based on patient and disease characteristics will be performed on an ad hoc basis. Reviews of the data will include evaluations of potential prognostic and predictive disease characteristics with specific interest in how these prognostic factors impact the EASI score. If a potential prognostic factor merits further consideration the exploratory analyses will at least include an NRI logistic regression analysis of the EASI-50 after 16 weeks where treatment, country and the potential factor are included in the model.

## **7 Safety Analyses**

All analyses of safety including the extent of exposure to study medication will be performed using the safety population. Reports will be generated using the overall population, and some reports will be repeated using only patients enrolled in Japan.



## 7.1 Study Medication Exposure and Treatment Compliance

### 7.1.1 Extent of Exposure

Extent of exposure will be derived as:

$$\text{Extent of Exposure} = \text{Date of 1}^{\text{st}} \text{ dose of study medication} - \text{Date of last dose of study medication} + 1,$$

where study medication considers only the study tablets. Extent of exposure will be descriptively summarized for each treatment arm.

Patients in this trial will also be given triamcinolone 0.1% cream as a concomitant therapy. The use of the cream will commence at the time of the first study visit and will continue throughout the treatment period of the study. The amount of triamcinolone 0.1% cream used will be derived for both the baseline period and the treatment period. In both cases the usage will be normalized to a 30 day period, i.e.

$$\text{Triamcinolone average monthly usage} = \frac{\text{total weight (in grams) dispensed} - \text{total weight (in grams) returned}}{\text{duration (in days) of study period} / 30 \text{ days}}$$

The baseline period will commence on the date that the triamcinolone 0.1% cream is first dispensed and end on the day prior to the date of first dose of randomly assigned study treatment. The treatment period will commence on the date of first dose of randomly assigned study treatment and end on the date of last treatment period visit, i.e. usually the earlier of the date of the Week 16 clinic visit or the date of the early termination visit. For both study periods, triamcinolone average monthly usage will be descriptively summarized for each treatment arm using the overall and by-country populations. If warranted after review of the data, additional analyses of select efficacy variables that incorporate the triamcinolone average monthly usage as a covariate in the statistical model may be considered.

### 7.1.2 Treatment Compliance

Treatment compliance with the randomly assigned study medication will be evaluated at every clinic visit, so that the clinical site can reinforced the importance of treatment compliance with the patient. A patient will be considered significantly noncompliant if he or she misses more than 20% of the prescribed doses during the study, i.e. compliance < 80%. Similarly, a patient will be considered significantly noncompliant if he or she is judged by the investigator to have intentionally or repeatedly taken more than the prescribed amount of medication, i.e. compliance  $\geq$  120%. Persistent non-compliance can result in the patient being discontinued from the study.

For the purpose of analysis only the overall treatment compliance will be summarized. Treatment compliance will be derived as

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

The total number of tablets expected will be equal to the number of days between the date of first dose of randomly assigned study treatment and the date of the patient's last treatment period visit.

Treatment compliance will be descriptively summarized for each treatment arm. In addition, the number and percentage of patients whose compliance < 80% or  $\geq$  120% will also be presented

for each treatment arm. Treatment compliance summaries will be produced using the ITT population.

## 7.2 Adverse Events

AEs will be collected from the day the patient gives informed consent until 30 days after the final dose of study drug or until resolution of all serious adverse events (SAEs). Only those AEs that emerge or worsen after the date of first dose of randomly assigned study medication, i.e. treatment emergent adverse events (TEAE), will be reported. Adverse events with a missing start date will also be classified as TEAE.

Summaries of TEAE will include the number of patients with at least one TEAE for each treatment arm. When reporting by system organ class (SOC) and preferred term (PT), the reports will present the SOC in alphabetical order; while PTs within the SOC will be presented in order of overall decreasing frequency of occurrence in the combined baricitinib arms. A patient with multiple TEAEs (different PTs) coded to the same SOC will be counted only once for that SOC, but will be counted each time for different PTs within that SOC. A patient with separate events of the same PT (different start/stop dates) will be counted only once in the frequency tables for that PT.

TEAEs will be reported using the treatment that the patient received.

No statistical testing will be performed for comparisons of TEAEs.

An overview of all TEAEs will also be provided by treatment group using the types of AEs defined in the following subsections.

### 7.2.1 Treatment-Emergent Adverse Events

TEAEs will be summarized by SOC and PT for each treatment arm. In addition, a presentation of TEAE preferred terms in decreasing frequency in the combined baricitinib treatment arms will be provided.

### 7.2.2 Treatment Related TEAE

Every AE will be assessed by the investigator for its relationship to the randomly assigned study medication. The subset of TEAE considered by the investigator as related to study treatment will be called treatment related TEAE.

Treatment related TEAE will be summarized for each treatment arm by SOC and PT and by decreasing frequency in the combined baricitinib arms.

### 7.2.3 Serious Treatment-Emergent Adverse Events

An SAE is an AE that met one or more of the following criteria:

- death
- initial or prolonged inpatient hospitalization
- a life-threatening experience (that is, immediate risk of dying)

- persistent or significant disability/incapacity
- congenital anomaly/birth defect or
- considered significant by the investigator for any other reason

Serious TEAEs will be summarized for each treatment arm by SOC and. A listing of all SAE will also be provided. The listing will be sorted by treatment arm, country, patient ID and study day of onset.

#### ***7.2.4 TEAE Resulting in Death***

If there are any TEAE that result in death, then a listing of all deaths will be provided. The listing will be sorted by treatment arm, country, patient ID and study day of onset.

#### ***7.2.5 TEAE Leading to Study Drug Discontinuation***

For every AE in the eCRF the investigator indicates whether the action taken with respect to the study medication was dose not changed or drug withdrawn.

The TEAE that lead to study drug discontinuation will be summarized for each treatment arm by SOC and PT and by decreasing frequency in the combined baricitinib. A listing of the TEAE that lead to study drug discontinuation will also be provided.

#### ***7.2.6 TEAE by Maximal Severity***

Every AE will be graded by the investigator as mild, moderate, or severe, so for each patient the greatest severity observed can be obtained by comparing the severity of all of a patient's TEAE that share the same SOC or PT. A table of TEAE by maximal severity will be prepared for each treatment arm by SOC and PT.

#### ***7.2.7 Adverse Events of Special Interest***

Adverse events or laboratory results of special interest include infections; myelosuppressive events of anaemia (haemoglobin decreased), leukopenia (white blood cell count decreased), neutropenia (neutrophil count decreased), lymphopenia (lymphocyte count decreased), thrombocytopenia (red blood cell count decreased), thrombocythaemia (platelet count increased); hyperlipidaemia (lipids increased); hypercholesterolemia (cholesterol increased); and elevations in hypertransaminasaemia (aminotransferase increased) and hyperbilirubinemia (bilirubin increased). All TEAE for which the PT contains any of the above terms or that belong to the SOC 'Infections and infestations' will be considered an AE of special interest. The TEAE of special interest will be summarized for each treatment arm by SOC and PT and by decreasing frequency in the combined baricitinib arms.

### **7.3 Clinical Laboratory Results**

Blood samples for hematology and serum chemistry and urine samples for urinalysis will be collected at baseline and Weeks 4, 8, 12, and 16 post-randomization. Blood samples for fasting lipid profiles will be collected at baseline and Week 16 post-randomization, and blood samples for immunoglobulins will be collected at baseline and Weeks 8 and 16 post-randomization. If a

patient decides to discontinue early from the study, samples for each of the parameters will be collected prior to discontinuation.

Analyses of laboratory values will be produced using the measurements collected the scheduled time points. The unscheduled assessments will only appear in data listings. An exception will be made for the analyses of the last measured values which for patients who do not have a measurement at Week 16 will be derived using the last collected scheduled or unscheduled measurement.

### ***7.3.1 Descriptive Statistics over Time***

At each of the post-baseline visits the change and percent change from baseline for each every quantitative clinical laboratory parameter will be derived. Descriptive statistics for the value and changes from baseline will be tabulated for each time point for each treatment arm.

### ***7.3.2 Shift tables and Categorical variables***

The grading of each laboratory parameter as low, normal or high will be compared from baseline to Week 16 and to the last observation, which should be the Week 16 or early termination visit. Shift tables will present the number and percent of patients who started in a category (low, normal, high) at baseline and category at the end of the study. Similarly, shift tables for the categorical urinalysis parameter will also be produced.

Shift tables comparing baseline to the other time points in the study may be generated if warranted after data review.

Additional shift tables will be constructed for the liver parameters – alanine aminotransferase (ALT), aspartate aminotransferase (AST), alkaline phosphatase (ALP), and total bilirubin (TBili). For each of the transaminases the number of patients with any post-baseline elevation exceeding 3 times the upper limit of the normal range (3xULN) will be counted by treatment arm. The counts of patients will be reported using subsets defined by the patients' baseline transaminase values; i.e.  $ALT \leq 1xULN$ ,  $1xULN < ALT \leq 2xULN$ . The analyses of ALT and AST elevations will also be performed using 5xULN and 10xULN. For TBili the number of patients with any post-baseline elevation exceeding 2xULN will be counted by treatment arm using the subsets  $TBili \leq 1xULN$  and  $TBili > 1xULN$ . For ALP the number of patients with any post-baseline elevation exceeding 1.5xULN (also 2xULN and 3xULN) will be counted by treatment arm.

## **7.4 Vital Signs and Weight**

The vital signs pulse and blood pressure will be assessed at baseline and at Weeks 1, 4, 8, 12, and 16 post-randomization, plus at the early termination visit, if applicable. Except for the Week 1 visit, weight will also be collected. At each of the post-baseline visits the vital signs and weight change from baseline will be computed. Descriptive statistics for the value and change from baseline will be tabulated for each time point. In addition, the number and percentage of patients with a systolic blood pressure reading either  $\leq 90$  mmHg or  $\geq 160$  mmHg or a change  $\geq 20$  mmHg, with a diastolic blood pressure reading either  $\leq 50$  mmHg or  $\geq 100$  mmHg or a change  $\geq 10$  mmHg, or with a pulse either  $< 50$  bpm or  $> 100$  bpm or a change  $\geq 15$  bpm will also be presented by treatment group at each time point.

No statistical analyses are planned. Any clinically significant findings should result in a diagnosis to be reported as an AE. If warranted by additional review of the data further analyses will be performed.

## **7.5 Biomarker Results**

Skin biopsies, serum, plasma, and whole blood ribonucleic acid (RNA) samples for non-genetic biomarker research will be collected where local regulations and external review boards allow. The analyses of biomarkers will be presented for the overall population and by country. Analyses of biomarkers are exploratory objectives, so additional post-hoc analyses may be added if review of the data indicates such analyses are warranted.

### ***7.5.1 Skin biopsies***

Skin biopsy samples from atopic dermatitis lesions and adjacent normal tissue will be collected to study biomarkers related to atopic dermatitis and/or related to the mechanism of action of baricitinib. Samples will be collected at baseline and at Week 16. An additional sample from the lesion will be collected at Week 4. The tissue biomarkers to be collected include IL-4, Ki67, K16, and epidermal thickness. Change and percentage change from baseline for each parameter will be derived for each patient at each post-baseline visit for each type of tissue for each patient. Descriptive statistics for the value, change and percentage change will be tabulated for each time point.

### ***7.5.2 Blood biomarkers***

Peripheral blood and serum samples will be collected at baseline and at Weeks 4 and 16. The blood biomarkers to be collected include eosinophils, IL-4, IL-5, IL-10, IL-12, IL-13, IL-19, IL-31, CCL17/TARC, IgE, lactate dehydrogenase (LDH), and interferon gamma (IFH-g). Change and percentage change from baseline for each parameter will be derived for each patient at each post-baseline visit for each type of tissue for each patient. Descriptive statistics for the value, change and percentage change will be tabulated for each time point. Descriptive statistics for the maximal post-baseline change and percent change will also be presented.

### ***7.5.3 Filaggrin genotyping***

A blood sample for genotyping of filaggrin will be collected at baseline (Visit 2). The percent wild type will be descriptively summarized for the overall population. Scatterplots of filaggrin genotype with the change and percent change in EASI after 16 weeks of treatment will be presented along with the associated Pearson's correlation coefficient will be provided for the overall population. In addition, for each treatment group side-by-side box plots of the filaggrin genotype will be produced for the patients who did/did not achieve EASI-50 after 16 weeks of treatment.

### ***7.5.4 Correlation between Baseline Biomarkers***

Pearson correlation coefficients will be calculated for each pair of biomarkers. The baseline values of these biomarkers will be used for this analysis, and all patients with baseline biomarker

data will be included. A table of all pairwise correlations will be generated for the overall population. The pairwise scatterplots may also be produced if warranted based on review of the data and results.

### ***7.5.5 Biomarkers as Measures of Disease Severity***

Pearson correlation coefficients will be calculated for each biomarker and efficacy variable pairing. The baseline values of the biomarkers (skin biopsy, blood biomarker, and filaggrin genotype) and the efficacy variables (EASI, SCORAD, IGA, DLQI, Itch NRS, and POEM) will be used for this analysis. All patients with baseline biomarker data and efficacy data will be included. A table of all pairwise correlations will be generated for the overall population. The pairwise scatterplots may also be produced if warranted based on review of the data and results.

## **8 Pharmacokinetic Analyses**

Venous blood samples (approximately 5 mL per sample) for the measurement of baricitinib concentrations will be drawn at baseline (pre-dose and at 15-30 minutes post-dose) and Weeks 4 (1.5-4 hours post-dose), 8 (4-8 hours post-dose), 12 (pre-dose), and 16 (30-90 minutes post-dose) post-randomization, plus at the early termination visit, if applicable. Plasma samples from patients on active treatment will be assayed for baricitinib using liquid chromatography and tandem mass spectrometry (LC/MS/MS).

Baricitinib concentrations will be summarized by treatment and time window using summary statistics (n, mean, standard deviation, coefficient of variation [CV], minimum, median, maximum, geometric mean, and geometric CV).

A population PK analysis will be performed separately and reported in a population PK report. Exposure-response relationships between the primary and select secondary efficacy endpoints and plasma exposure may also be explored.

## **9 Health Outcome Psychometric Analyses**

A separate analysis plan (GPORWE 2016-0067) describing the psychometric analyses of the Itch NRS and POEM as well as responder definition will be provided by Lilly Global Patient Outcomes and Real World Evidence.

## References

- 1 Gillings D, Koch G. The application of the principle of intention to treat to the analysis of clinical trial. *Drug Inf J*. 1991;25:411–424.