I4V-MC-JAIV Statistical Analysis Plan

I4V-MC-JAIV: A Randomized, Double-Blind, Placebo-Controlled, Proof-of-Concept Study Evaluating the Efficacy and Safety of Baricitinib (LY3009104) in Patients with Primary Biliary Cholangitis Who Have an Inadequate Response or Are Intolerant to UDCA

NCT03742973

Approval Date: 14-May-2019

1. Statistical Analysis Plan: I4V-MC-JAIV: A Randomized, Double-Blind, Placebo-Controlled, Proof-of-Concept Study Evaluating the Efficacy and Safety of Baricitinib (LY3009104) in Patients with Primary Biliary Cholangitis Who Have an Inadequate Response or Are Intolerant to UDCA

Baricitinib (LY3009104) Primary Biliary Cholangitis

Study I4V-MC-JAIV (JAIV) is a proof-of-concept, Phase II, multicenter, randomized, double-blind, parallel, placebo-controlled study to evaluate the safety and efficacy of baricitinib 2-mg and 4-mg compared to placebo in adult patients with primary biliary cholangitis (PBC) who have an inadequate response or are intolerant to UDCA.

Eli Lilly and Company
Indianapolis, Indiana USA 46285

Protocol I4V-MC-JAIV
Phase II

Statistical Analysis Plan electronically signed and approved by Lilly on date provided below.

Approval Date: 14-May-2019 GMT

2. Table of Contents

Section **Page**

Table of Contents

Table **Page**

Table of Contents

Figure **Page**

3. Revision History

Statistical Analysis Plan (SAP) Version 1 is based on Protocol I4V-MC-JAIV(b) and was approved prior to unblinding.

4. Study Objectives

4.1. Primary Objective

The primary objective of Study I4V-MC-JAIV (JAIV) is:

Table JAIV.4.1. Primary Objective and Endpoint

Objectives	Endpoints
• To evaluate the effect of baricitinib 4-mg QD compared to placebo on PBC disease	• Change from baseline in ALP at Week 12

Abbreviations: ALP = alkaline phosphatase; PBC = primary biliary cholangitis; QD = once daily.

According to ICH E9R1 guidelines, the objectives of a study should be translated into precise definition of estimands. Estimands are defined using the following four attributes:

- The population of interest consists of the patients targeted by the scientific question;
- The variable (or endpoint) of evaluation that is obtained for each patient, required to address the scientific question;
- The population-level summary of variable, a basis for treatment evaluation;
- The specification of how to account for intercurrent events to reflect the scientific question of interest.

In particular, for Study JAIV, the estimands are defined by:

- Population: The population of interest is patients with primary biliary cholangitis (PBC) who have had an inadequate response or intolerance to ursodeoxycholic acid (UDCA). Analysis populations are defined in Section 6.1.1. The analysis population corresponding to each of the efficacy and health outcome endpoints is specified in Table JAIV.6.3 and Table JAIV.6.5.
- Variables: The primary endpoint/variable is listed in Table JAIV.4.1 and secondary endpoints/variables are listed in Table JAIV.4.2. A full list of efficacy and health outcome endpoints/variables is given in Table JAIV.6.2 and Table JAIV.6.4.
- Population-level Summary: The summary measure for binary variables will be proportion and the summary measure for continuous variables will be average. The primary comparison of interest will be the difference of proportions between baricitinib dose groups and placebo. Details are given in Section 6.7.
- Intercurrent Event(s) Strategy: Intercurrent events are defined as certain events occurring during the course of the study which complicate the interpretation of the treatment effects, such as treatment discontinuation and death. Intercurrent events will be handled by the while-on-treatment strategy and the hypothetical strategy approaches as primary strategies. For the while-on treatment strategy, the variable of interest is the response prior to the occurrence of the event, i.e., while the subject was on treatment, whereas the

hypothetical strategy aims to estimate the effect in the hypothetical situation that all the patients continue in the study to the time point of interest and adhere to their randomized treatment for the entire time. The composite strategy, which incorporates the occurrence of an intercurrent event as a component of the response of interest, will also be used in supplementary analysis. An example of composite strategy might be if early treatment discontinuation is considered a form of treatment failure. Specific statistical methods to be used for handling intercurrent events under different strategies are described in Section 6.3. Statistical methods corresponding to each of the efficacy and health outcome variables are summarized in Table JAIV.6.2 and Table JAIV.6.4.

4.2. Secondary Objectives

The secondary objectives of the study are the following:

Table JAIV.4.2. Secondary Objectives and Endpoints

Objectives	Endpoints
• To evaluate the effect of baricitinib 4-mg QD compared to placebo on PBC disease	• Proportion of patients with ALP <1.67 x ULN (and at least 15% decrease from baseline) and total bilirubin<ULN at Week 12
• To evaluate the effect of baricitinib 2-mg QD compared to placebo on PBC disease	• Change from baseline in ALP at Week 12 • Proportion of patients with ALP<1.67 x ULN (and at least 15% decrease from baseline) and total bilirubin < ULN at Week 12
• To evaluate the effect of baricitinib 4-mg and 2-mg QD compared to placebo on PBC symptoms	• Change from baseline in itch as measured by Itch NRS at Week 12 • Change from baseline in fatigue as measured by Fatigue NRS at Week 12

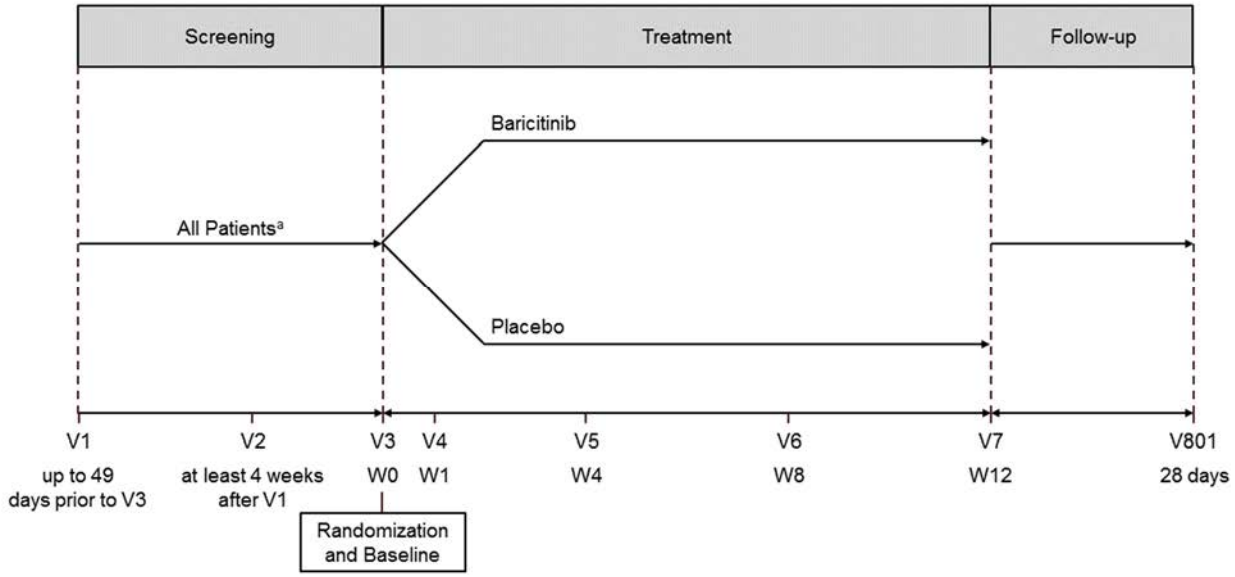
Abbreviations: ALP = alkaline phosphatase; NRS = Numeric Rating Scale; PBC = primary biliary cholangitis; QD = once daily; ULN = upper limit of normal.

5. Study Design

5.1. Summary of Study Design

Study JAIV is a multicenter, randomized, double-blind, parallel, placebo-controlled, proof-of-concept, Phase II study in patients with primary biliary cholangitis (PBC) who have an inadequate response or are intolerant to UDCA. The study will be conducted in 2 cohorts. Cohort A will randomize patients to placebo and baricitinib 2-mg. Cohort B will randomize patients to placebo and baricitinib 4-mg. Enrollment in Cohort B will begin only after the data monitoring committee (DMC) review of Cohort A is complete and Lilly has received and accepted a recommendation to proceed. Both cohorts will follow the same study schedule.

Figure JAIV.5.1 illustrates the study design.



Abbreviations: V = Visit; W = Week.

[a] Patients able to tolerate ursodeoxycholic acid (UDCA) will take UDCA orally as background therapy during the study.

Figure JAIV.5.1. Illustration of study design for Clinical Protocol I4V-MC-JAIV. Patients in Cohort A and Cohort B will follow the same study schedule as outlined in this figure. Cohort A will randomize 12 patients to baricitinib 2-mg and 8 patients to placebo. After a DMC review of Cohort A data, Cohort B will then randomize 20 patients to baricitinib 4-mg and 12 patients to placebo.

5.2. Determination of Sample Size

Approximately 20 patients will be randomized to the baricitinib 2-mg and placebo treatment groups for Cohort A. Twelve patients will be randomly assigned to the 2-mg treatment group and 8 patients to the placebo treatment group. This sample size is appropriate to provide

sufficient preliminary safety and tolerability information for baricitinib 2-mg in patients with PBC prior to studying 4-mg baricitinib.

For Cohort B, 32 patients will be randomized to the baricitinib 4-mg and placebo treatment groups. Twenty patients will be randomized to the baricitinib 4-mg treatment group and 12 patients to placebo.

Overall, 20 patients per group will be assigned to the baricitinib 4-mg (Cohort B) and placebo (Cohorts A and B) dose groups. Based on this sample size, this study will have approximately 81% power to detect a difference in ALP change from baseline between baricitinib 4-mg and placebo at Week 12 assuming that the true difference is 65 U/L. The sample size was determined based on a 2-sample t-test with 2-sided alpha level of 0.1 and a standard deviation of 80 U/L for the change from baseline in ALP. The sample size calculation was performed using nQuery® Advisor 7.0 software.

5.3. Method of Assignment to Treatment

[redacted]

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 double-blind investigational product to each patient. Site personnel will confirm that they have located the correct package by entering a confirmation number found on the package into the IWRS.

6. A Priori Statistical Methods

6.1. General Considerations

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

Statistical analysis of this study will be the responsibility of Eli Lilly and Company (Lilly) or its designee. The statistical analyses will be performed using SAS® Version 9.4 or a more recent version.

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 would be available upon request.

Incidence rates and 95% confidence interval (CI) will be displayed for select safety endpoints.

Statistical tests of treatment effects and CIs will be performed at a 2-sided significance level of 0.1, unless otherwise stated. No multiplicity adjustment will be used in this study.

Data collected at early termination visits will be mapped to the next scheduled visit number for that patient if it falls within the visit window. 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 be included in other analyses, including shift analyses for safety analytes, change from baseline to endpoint using last observation carried forward (LOCF) analyses, and other categorical analyses.

6.1.1. Analysis Populations

Modified intent-to-treat (mITT) population: The mITT population is defined as all randomized patients who take at least 1 dose of investigational product

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

Unless otherwise specified, the efficacy and health outcome analyses will be conducted on the mITT population. Patients will be analyzed according to the treatment to which they were randomized. The analysis populations will include patients from both Cohort A and Cohort B.

Concomitant medications will be descriptively summarized by treatment group in terms of frequencies and percentages using the safety population. The medications will be coded accordingly.

All safety data will also be descriptively summarized by treatment groups and analyzed using the safety population.

In the rare situation where a patient is lost to follow-up at the first postbaseline visit, but some safety data exist (e.g., unscheduled laboratory assessments) after first dose of study drug, a listing of the data or a patient profile will be provided.

6.1.2. Definition of Baseline and Postbaseline Measures

Baseline will be defined as the last available value before the first dose of investigational product for both efficacy and safety analyses. In most cases, this will be the measure recorded at Week 0 (Visit 3). Change from baseline will be calculated as the visit value of interest minus the baseline value.

Postbaseline measurements are collected after study drug administration through Visit 7 (Week 12) or an early discontinuation visit.

6.1.3. Analysis Methods

The primary analysis for all continuous efficacy and health outcome variables will use a mixed-effects models for repeated measures (MMRM) with [redacted] An unstructured covariance matrix will be used to model the within-patient variance-covariance errors. If the unstructured covariance matrix results in a lack of convergence, the heterogeneous Toeplitz covariance structure, followed by the heterogeneous autoregressive covariance structure will be used. The Kenward-Roger method will be used to estimate the denominator degrees of freedom. Type III tests will be used for all the statistical comparisons; the 90% confidence interval (CI) will also be reported. The LS mean difference, standard error, p-value, and CI will also be reported. Treatment group comparisons at specific study visits will be tested using the t-test obtained from the MMRM results. Data will be censored as described in Section 6.3.

The primary analysis for all categorical efficacy and health outcome variables will use a logistic regression analysis [redacted] The p-value and 90% (unless otherwise specified) CI for the odds ratio from the logistic regression model will be used for primary statistical inference. When logistic regression sample size requirements are not met, (<5 responders in any category for any factor), the p-value from the Cochran-Mantel-Haenszel (CMH) test with [redacted]

In addition, all primary and secondary continuous efficacy and health outcome variables will use analysis of covariance (ANCOVA) models [redacted] and cohort in the model as a supplementary analysis. Type III sums of squares for the least-squares mean (LSM) will be used for the statistical comparison of treatment groups, and the LSM difference, standard error, p-value and 90% CI will also be reported. The last observation carried forward (LOCF) approach will be used to impute missing data.

All safety data will be descriptively summarized by treatment groups and analyzed using the safety population.

Treatment-emergent adverse event (TEAEs) are defined as AEs that first occurred or worsened in severity after the first dose of study treatment. The number of TEAEs as well as the number and percentage of patients who experienced at least 1 TEAE will be summarized using the Medical Dictionary for Regulatory Activities (MedDRA) for each system organ class (or a body system) and each preferred term by treatment group. SAEs and AEs that lead to discontinuation of investigational product and discontinuation from the study will also be summarized by treatment group. The Fisher exact test will be used to perform comparisons between the baricitinib and placebo groups.

All clinical laboratory results will be descriptively summarized by treatment group. Individual results that are outside of normal reference ranges will be flagged in data listings. Quantitative clinical hematology, chemistry, and urinalysis variables obtained at the baseline and postbaseline visits will be summarized as changes from baseline by treatment group and analyzed using analysis of covariance (ANCOVA) with treatment 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 least-squares mean (LSM) changes from baseline are different from zero; the p-value and standard error for the LSM change will 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. Treatment-emergent high/low for categorical laboratory safety analyses will also be produced. Categorical variables, including the incidence of abnormal values and incidence of AEs of special interest, will be summarized by frequency and percentage of patients in corresponding categories. Shift tables will be presented for selected measures.

Observed values and changes from baseline (predose or screening if predose is missing) for vital signs and physical characteristics will be descriptively summarized by treatment group and visit. Change from baseline to postbaseline in vital signs and body weight will be analyzed using ANCOVA with treatment and baseline value in the model.

The incidence and average duration of investigational product interruptions will be summarized and compared descriptively among treatment groups.

6.2. Covariate Adjustment

The randomization to treatment groups at Week 0 (Visit 3) is stratified by [redacted]

The covariates used in the ANCOVA model for continuous data generally will include the parameter value at baseline. Inclusion of baseline in the ANCOVA model ensures that 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.

6.3. Handling of Dropouts or Missing Data

Last-observed-value-carried forward (LOCF) (for categorical variables) and MMRM (for continuous variables) will be the primary methods used to handle missing data. Nonresponder imputation (NRI) will also be used as a supplementary analysis method for categorical data. The censoring rules along with their associated estimator assumptions are described in Sections 6.3.1 through 6.3.3.

As efficacy and health outcome data can accrue after a patient permanently discontinues study drug, general censoring rules to the data will be applied to all efficacy and health outcome observations subsequent to these events.

The censoring rule will censor efficacy and health outcome results after permanent study drug discontinuation. This censoring rule will be applied to all continuous and categorical efficacy and health outcome endpoints. Sections 6.3.1 through 6.3.3 summarize the imputation methods for the various efficacy and health outcome endpoints.

6.3.1. Last Observation Carried Forward

Categorical data will primarily be imputed by carrying forward the last post-baseline assessment for the measure. According to ICH E9 R1, the while-on-treatment strategy could be applied based on the last post-baseline value (scheduled or unscheduled visits) at or before the visit of interest while the patient was still on study drug. Patients in the analysis population without at least 1 postbaseline observation will be defined as non-responders for this analysis. After last observation carried forward (LOCF) imputation is applied, data from patients with nonmissing baseline and at least 1 postbaseline observation will be included in the analyses. These LOCF analyses help ensure the maximum number of randomized patients who were assessed post-baseline will be included in the analyses.

All categorical endpoints will utilize LOCF methodology after censoring efficacy and health outcome results for those who permanently discontinued study drug. For all patients who permanently discontinue study treatment or discontinue from the study for any reason at any time, the last nonmissing postbaseline observation on or prior to discontinuation will be carried forward to subsequent time points for evaluation. If a patient does not have a nonmissing observed record (or one imputed by other means) for a postbaseline visit prior to discontinuation, the last postbaseline record prior to the missed visit will be used for the visit.

Further, all continuous primary and secondary efficacy and health outcome variables will be imputed using LOCF as a supplementary analysis.

6.3.2. Mixed Model for Repeated Measures

For the continuous primary, secondary, and exploratory efficacy and health outcome variables, such as ALP and Itch NRS, MMRM analyses will be performed to mitigate the impact of missing data. This approach assumes that missing observations are missing-at-random (missingness is related only to observed data and not to any unobserved data) during the study and takes into account both the missingness of data and the correlation of the repeated

measurements. An MMRM method can be justified based on the hypothetical strategy (ICH E9 R1) for handling intercurrent events.

All continuous endpoints will utilize MMRM after censoring efficacy and health outcome results for those who permanently discontinued study drug.

6.3.3. Nonresponder Imputation

For the analysis of all categorical efficacy and health outcome variables, the nonresponder imputation (NRI) method will be used as a supplementary analysis. NRI can be justified based on the composite strategy (ICH E9R1) for handling intercurrent events. In this strategy a patient is defined as a responder only if (i) they meet the clinical requirements for response at the predefined time and (ii) they remain on the assigned study treatment. Failing either criteria by definition makes them a non-responder.

All secondary categorical endpoints will utilize the NRI method after censoring efficacy and health outcome results for those who permanently discontinued study drug. Patients in the analysis population without at least 1 post-baseline observation will also be defined as nonresponders for all visits. As well, patients who are missing a value prior to discontinuation (ie, the patient is missing an intermediate visit) will be imputed as nonresponders at that visit.

6.4. Multicenter Studies

This study will be conducted by multiple investigators at multiple sites internationally. However, since the number of patients at each site is expected to be very small, no further sensitivity analysis will be performed for the sites or countries.

6.5. Multiple Comparisons/Multiplicity

No multiplicity adjustment will be used in this study.

6.6. Patient Disposition

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

Patient disposition through Week 12 will be summarized using the mITT population. Frequency counts and percentages of patients who complete the study treatment visits or discontinue early from the study along with whether they completed follow-up or did not complete follow-up will be summarized separately by treatment group for patients along with their reason for study discontinuation. Frequency counts and percentages of patients who complete the treatment or discontinue treatment early will also be summarized separately by treatment group for patients along with their reason for treatment discontinuation.

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

6.7. Patient Characteristics

Patient characteristics including demographics and baseline characteristics will be summarized descriptively by treatment group for the mITT population. Historical illnesses and pre-existing conditions will be summarized descriptively by treatment group for the safety population. Descriptive statistics including number of patients (n), mean, SD, minimum, 1st quartile, median, 3rd quartile, and maximum will be provided for continuous measures, and frequency counts and percentages will be tabulated for categorical measures. A listing of patient demographics will also be provided for the mITT population. No formal statistical comparisons will be made among treatment groups unless otherwise stated.

Table JAIV.6.1 describes the specific variables and how they will be summarized.

Table JAIV.6.1. Patient Characteristics

Variable	Continuous measure Summary	Categorical Summary
Age [a]	Yes	None
Sex	No	Male, Female
Race	No	American Indian/Alaska Native, Asian, Black/African American, Native Hawaiian or other Pacific Islander, White, or Multiple
Geographic region	No	By Country
Height (cm)	Yes	None
Weight (kg)	Yes	None
BMI[b]	Yes	<18.5 kg/m^2, ≥18.5 and <25 kg/m^2, ≥25 and <30 kg/m^2, ≥30 and <35 kg/m^2 , ≥35 and <40 kg/m^2, ≥40 kg/m^2
Alcohol use	No	Never, Current, Former
Tobacco use	No	Never, Current, Former
Age at PBC diagnosis (years)[c]	Yes	<18 years, ≥18 years
Duration of PBC diagnosis[d]	Yes	0 to <2 years, 2 to <5 years, 5 to <10 years, 10 to <20 years, ≥20 years
Current UDCA use	No	Yes, No
Baseline ALP level	Yes	Baseline ALP ≤2.5xULN, Baseline ALP >2.5xULN
Quality of Life for Primary Biliary Cirrhosis (PBC-40) at baseline	Yes	None
Functional Assessment of Chronic Illness Therapy-Fatigue (FACIT-F) at baseline	Yes	None
Itch Numeric Rating Scale (NRS) at baseline	Yes	No pruritus (0), Mild pruritus (>0 to <4), Moderate pruritus (≥4 to <7), Severe pruritus (≥7 to <9), Very severe pruritus (≥9)
		Frequency table for each value
Fatigue NRS	Yes	None
Physician's Global Assessment of Disease Activity at baseline	Yes	None
Total bilirubin at baseline	Yes	None

Abbreviations: ALP = alkaline phosphatase; BMI = body mass index; NRS = Numeric Rating Scale; PBC = primary biliary cholangitis; UDCA = ursodeoxycholic acid.

[a] Age in years will be calculated as (number of months between first dose date and July 1st of birth year) / 12, and truncated to a whole-year (integer) age.

[b] Body Mass Index (BMI) will be calculated as: $BMI\ (kg\ /\ m^2) = Weight\ (kg)\ /(Height\ (m))^2$.

[c] Age at diagnosis in years will be calculated as (number of months between date of PBC diagnosis and July 1st of birth year) / 12, and truncated to a whole-integer age.

[d] The duration of PBC from diagnosis (year) = [(Date of informed consent – Date of PBC diagnosis)+1]/ 365.25. If year of onset is missing, duration of PBC will be set as missing. Otherwise, unknown month will be taken as January, and unknown day will be taken as 01. The duration of PBC will be rounded to 1 decimal place.

6.7.1. 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 safety 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, the Prespecified Medical History: Comorbidities eCRF, or the Adverse Events eCRF with a start date prior to the date of informed consent and an end date on or after informed consent or no end date (that is, the event is ongoing). For events occurring on the day of the first dose of study treatment, the date and time of the onset of the event will both be used to determine if the event was pre-existing. Conditions with a partial or missing start date (or time if needed) will be assumed to be "not pre-existing" unless there is evidence, through comparison of partial dates, to suggest otherwise. Pre-existing conditions will be categorized using the MedDRA SMQs or similar pre-defined lists of PTs of interest. Frequency counts and percentages of patients with selected pre-existing conditions will be summarized by treatment group using the safety population. If a preexisting condition worsens in severity on or after date of first dose of study treatment, it will be considered a TEAE from the date of worsening onwards.

6.8. Treatment Compliance

Compliance with investigational product treatment for baricitinib and placebo will be assessed through counts of returned investigational product tablets. A patient will be considered significantly noncompliant if he or she misses more than 20% of the prescribed doses during the study, unless the patient's investigational product was withheld by the investigator for safety reasons.

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.

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

where

- Total number of tablets dispensed: sum of tablets dispensed in the period of interest prior to Visit *x*;
- Total number of tablets returned: sum of the tablets returned in the period of interest prior to and including Visit *x*;

- Expected number of tablets: number of days in the period of interest*number of tablets taken per day = [(date of visit – date of first dose + 1) – number of days of temporary drug interruption]*number of tablets taken per day

Descriptive statistics for percent compliance and non-compliance rate will be summarized for the mITT population by treatment group for Week 0 through Week 12. 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 12.

6.9. Concomitant and Previous Therapy

Concomitant medications will be coded and descriptively summarized by treatment group in terms of frequencies and percentages using the mITT population.

At screening, previous and current PBC treatments are recorded for each patient. A summary of previous medications used for PBC, including zoster immunization 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 4-mg group. Concomitant therapy will be recorded at each visit and will be classified similarly. An additional summary for previous medications used for PBC will be created containing the reason for discontinuation.

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

Summaries of previous medications will be provided for the following categories:

- Previous PBC therapies
- Previous PBC therapies including reason for discontinuation

Summaries of concomitant medications will be provided for the following categories:

- Concomitant medications for PBC
- Concomitant medications for non-PBC

6.10. Efficacy Analyses

The general methods used to summarize efficacy data, including the definition of baseline value for assessments are described in Section 6.1.

Efficacy will generally be analyzed from Week 0 to Week 12, and patients will be analyzed according to the investigational product to which they were randomized at Week 0 (Visit 3):

Table JAIV.6.2 includes the descriptions and derivations of the primary and secondary efficacy outcomes.

Table JAIV.6.3 provides the detailed analyses including analysis type, method and imputation, population, time point, and comparisons for efficacy analyses.

Table JAIV.6.2. Description and Derivation of Primary and Secondary Efficacy Outcomes

Measure	Description	Variable	Derivation / Comment	Definition of Missing
Alkaline Phosphatase (ALP)	Laboratory measurement of the enzymatic (catalytic) activity of ALP enzyme within the blood.	▪ Change from baseline in ALP	Change from baseline: observed ALP – baseline ALP	Missing if baseline or observed value is missing.
Total Bilirubin and Alkaline Phosphatase (ALP)	Measurement of concentration of total bilirubin in the blood and measurement of the enzymatic (catalytic) activity of ALP enzyme within the blood.	▪ Proportion of patients with ALP <1.67 x ULN (and at least 15% decrease from baseline) and total bilirubin < ULN	If observed ALP < 1.67 x ULN and $\frac{baseline\ ALP - observed\ ALP}{baseline\ ALP} \times 100 \geq 15$, and observed total bilirubin <ULN then conclude that the patient has ALP <1.67 x ULN (and at least 15% decrease from baseline) and total bilirubin < ULN.	Missing if any one of baseline ALP, observed ALP, or observed total bilirubin is missing.

Table JAIV.6.3. Description of Primary and Secondary Efficacy Analyses

Measure	Variable	Analysis Method (Section 6.1.3)	Population (Section 6.1.1)	Comparison/Time Point	Analysis Type
Alkaline Phosphatase (ALP) and Total Bilirubin	• Change from baseline in ALP	MMRM	mITT	Bari 4-mg vs PBO; Week 12	Primary analysis
				Bari 2-mg vs PBO; Week 12	Secondary analysis
		ANCOVA using LOCF	mITT	Bari 4-mg vs PBO; Week 12	Supplementary analysis
				Bari 2-mg vs PBO; Week 12	Supplementary analysis
	• Proportion of patients with ALP <1.67 x ULN (and at least 15% decrease from baseline) and total bilirubin<ULN at Week 12	Logistic regression using LOCF	mITT	Bari 4-mg vs PBO; Week 12	Secondary analysis
				Bari 2-mg vs PBO; Week 12	Secondary analysis
		Logistic regression using NRI	mITT	Bari 4-mg vs PBO; Week 12	Supplementary analysis
				Bari 2-mg vs PBO; Week 12	Supplementary analysis

Abbreviations: ANCOVA = analysis of covariance; Bari = baricitinib; mITT = modified intent-to-treat; LOCF = last observation carried forward; MMRM = mixed model repeated measures; NRI = nonresponder imputation; PBO = placebo.

6.10.1. Primary Outcome and Methodology

The primary efficacy assessment in Study JAIV will be change in ALP at Week 12. The primary analysis of the study is to test the hypothesis that baricitinib 4-mg is superior to placebo for PBC disease based on the mITT population. MMRM described in Section 6.3.2 will be the primary method of analysis. The treatment difference along with the p-value and 90% confidence intervals will be reported.

6.10.2. Secondary Efficacy Analyses

There will be no adjustment for multiple comparisons for Study JAIV. The secondary efficacy analyses are detailed in Table JAIV.6.3. Health outcome analyses are described in Section 6.11.

6.10.3. Supplementary Analyses

Supplementary analyses for select outcomes have been previously described and include the following:

- Non-responder imputation (NRI) (Section 6.3.3) for categorical secondary outcomes.
- Analysis of continuous outcomes with ANCOVA (Section 6.1.3), with missing data imputed using LOCF (Section 6.3.1) for continuous primary and secondary outcomes.

6.11. Health Outcomes/Quality-of-Life Analyses

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

Health outcomes and quality-of-life measures will generally be analyzed according to the formats discussed in Section 6.10.

Table JAIV.6.4 includes the descriptions and derivations of the health outcomes and quality-of-life measures.

Table JAIV.6.5 provides the detailed analyses including analysis type, method and imputation, population, time point, and comparisons for health outcomes and quality-of-life measures.

Table JAIV.6.4. Description and Derivation of Health Outcomes and Quality-of-Life Measures

Measure	Description	Variable	Derivation / Comment	Definition of Missing
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 describes the worst level of itching in the past 7 days.	▪ Change from baseline in Itch NRS	Change from baseline: observed Itch score – baseline Itch score	Missing if baseline or observed value is missing.
Fatigue Numeric Rating Scale (NRS)	The Fatigue NRS is a single item, patient administered, 11 point horizontal scale anchored at 0 and 10, with 0 representing "no fatigue" and 10 representing "as bad as you can imagine." Overall severity of a patient's fatigue is indicated by selecting the number that describes the worst level of fatigue during the past 7 days.	▪ Change from baseline in Fatigue NRS	Change from baseline: observed Fatigue score – baseline Fatigue score	Missing if baseline or observed value is missing.
Functional Assessment of Chronic Illness Therapy Fatigue (FACIT-F)	The FACIT-F measures an individual's self-reported level of fatigue during their usual daily activities over the past 7 days. The scale is composed of 13 items measured on a 4-point Likert scale (4 = very much to 0 = not at all). The total score ranges from 0 to 52, and higher scores represent less fatigue. A score of less than 30 indicates severe fatigue (Webster et al. 2003).	▪ Change from baseline in FACIT-F total score	Change from baseline: observed FACIT-F total score – baseline FACIT-F total score according to version 4 of the scoring guidelines. (http://www.ser.es/wp-content/uploads/2015/03/FACIT-F_INDICE.pdf) Individual item scoring will be done for exploratory purposes.	Missing individual items within the questionnaire are acceptable as long as more than 50% of the items are answered, e.g. 7 of 13. If less than 7 items are answered baseline or observed value is missing.

Measure	Description	Variable	Derivation / Comment	Definition of Missing
Quality of Life for Primary Biliary Cirrhosis (PBC-40)	The PBC-40 is a disease-specific, 40-item, patient-reported survey containing 6 domains: overall symptoms, itch, fatigue, cognition, social, and emotional. Response options for overall symptoms, itch, fatigue, and cognition are on a 5-point scale: "Never," "Rarely," "Sometimes," "Most of the time," and "Always." Response options are also on a 5-point scale for social and emotional: "Not at all," "A little," "Somewhat," "Quite a bit," and "Very much." Items are scored from 1 to 5 and the individual item scores are summed to give a total domain score where high scores represent high impact of PBC on quality of life (Jacoby et al. 2005).	▪ Change from baseline in PBC-40 for each of the 6 domain scores	Change from baseline: observed PBC-40 score – baseline PBC-40 score for each of the 6 domains	Missing if baseline or observed value is missing.

Measure	Description	Variable	Derivation / Comment	Definition of Missing
Physician's Global Assessment of Disease Activity	The Physician's Global Assessment of Disease Activity is the physician's assessment of the patient's overall disease activity due to PBC, as compared with all possible patients with PBC. The Physician's Global Assessment of Disease Activity is scored using a 100-mm visual analog scale (VAS), where 0 mm (measured from the left starting point of the line) indicates no disease activity and 100 mm (measured from the left starting point of the line) indicates the most severe disease activity possible for all patients with PBC. The Physician's Global Assessment of Disease Activity score is indicated by making a vertical tick mark on the line between 0 and 100 mm. There are benchmarks of 0 (0 mm), 1 (33 mm), 2 (67 mm), and 3 (100 mm) on the line corresponding to no, mild, moderate, and severe PBC disease activity, respectively.	▪ Change from baseline in Physician's Global Assessment of Disease Activity	Change from baseline: observed Physician's Global Assessment of Disease Activity score – baseline Physician's Global Assessment of Disease Activity score	Missing if baseline or observed value is missing.

Table JAIV.6.5. Description of Health Outcomes and Quality-of-Life Measures Analyses

Measure	Variable	Analysis Method (Section 6.1.3)	Population (Section 6.1.1)	Comparison/Time Point	Analysis Type
Itch Numeric Rating Scale (NRS)	• Change from baseline in itch	MMRM	mITT	Bari 4-mg and Bari 2-mg vs PBO; Week 12	Secondary analysis
		ANCOVA using LOCF	mITT	Bari 4-mg and Bari 2-mg vs PBO; Week 12	Supplementary analysis
Fatigue Numeric Rating Scale (NRS)	• Change from baseline in fatigue	MMRM	mITT	Bari 4-mg and Bari 2-mg vs PBO; Week 12	Secondary analysis
		ANCOVA using LOCF	mITT	Bari 4-mg and Bari 2-mg vs PBO; Week 12	Supplementary analysis
Functional Assessment of Chronic Illness Therapy Fatigue (FACIT-F)	• Change from baseline in FACIT-F total score	MMRM	mITT	Bari 4-mg and Bari 2-mg vs PBO; Week 12	Exploratory analysis
Quality of Life for Primary Biliary Cirrhosis (PBC-40)	• Change from baseline in PBC-40 domain scores	MMRM	mITT	Bari 4-mg and Bari 2-mg vs PBO; Week 12	Exploratory analysis
Physician's Global Assessment of Disease Activity	• Change from baseline in Physician's Global Assessment of Disease Activity	MMRM	mITT	Bari 4-mg and Bari 2-mg vs PBO; Week 12	Exploratory analysis

Abbreviations: ANCOVA = analysis of covariance; Bari = baricitinib; mITT = modified intent-to-treat; LOCF = last observation carried forward; MMRM = mixed model repeated measures; NRI = nonresponder imputation; PBO = placebo.

6.12. Bioanalytical and Pharmacokinetic/Pharmacodynamic Methods

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

6.13. Safety Analyses

Analysis populations are defined in Section 6.1.1. Table JAIV.6.6 describes the analysis sets. "Analysis set" is a broad term used to define a set of definitions including the population, the treatment group(s), the time period(s), and the comparison(s) where applicable.

Table JAIV.6.6. Definition of Analysis Sets

Analysis Set	Analysis Population	Treatment Groups (Short Label)	Baseline Time Period	Postbaseline Time Period	Inferential Comparisons
Treatment (T)	Safety Population	Placebo, Bari 2 mg, Bari 4 mg	From screening to just before first study drug administration	From immediately after first dose of study drug to up to 30 days off drug follow-up time	When used: Bari 2 mg vs. placebo; Bari 4 mg vs. placebo
Treatment plus follow-up (T + FU)	Safety Population	Placebo, Bari 2 mg, Bari 4 mg	From screening to just before first study drug administration	From immediately after first dose of study drug to end of study participation.	When used: Bari 2 mg vs. placebo; Bari 4 mg vs. placebo

Abbreviation: Bari = baricitinib.

The planned safety analyses are consistent with compound-level safety standards, which are based on multiple sources, including company standards, internal and external subject matter experts, and cross-industry initiatives (e.g., white papers produced by a PhUSE Computational Science Working Group (a collaboration with FDA and PhUSE), published in the PhUSE Deliverables Catalog at https://www.phuse.eu/white-papers). Descriptions of the safety analyses are provided in this SAP; however, additional details (such as handling of unscheduled visits, missing data, etc.) are in compound-level safety standards.

The following statistical methods will be used for safety unless otherwise noted:

- The Fisher exact test will be used for treatment comparisons of proportions, and odds ratios with corresponding 95% confidence intervals will be presented where specified.
- Treatment differences in mean change for continuous measurements will be assessed using an analysis of covariance (ANCOVA) model fitting "baseline" as a covariate. Type 3 sums of squares will be used. In addition, the crude mean, standard deviation (SD), minimum, median, and maximum will be displayed.

Tests with two-sided p-values less than or equal to 0.05 will be referred to as having strong statistical evidence for a treatment difference, unless otherwise noted. However, p-values should not be over-interpreted, except for those associated with pre-specified hypotheses, since they correspond to data-driven hypotheses and hence are only useful as a flagging mechanism.

Additional safety analyses will be performed as part of the integrated safety analyses.

6.13.1. Extent of Exposure

Duration of exposure (in days) to study drug will be summarized for the safety population by treatment group using descriptive statistics (n, mean, SD, minimum, 1st quartile, median, 3rd quartile, maximum). Cumulative exposure and duration of exposure will be summarized in terms of frequency counts and percentages by category and treatment group.

Duration of exposure will be calculated as follows:

- Duration of exposure to investigational product: *date of last dose of treatment – date of first dose of study drug + 1*.

Last dose of treatment is calculated as last date on treatment.

Total patient-years (PY) of exposure will be reported for each treatment group for overall duration of exposure. Descriptive statistics (n, mean, SD, minimum, 1st quartile, median, 3rd quartile and maximum) will be provided for patient-days of exposure and the frequency of patients falling into different exposure ranges will be summarized. Exposure ranges will generally be reported in weeks using the following as a general guide:

- ≥4 weeks, ≥8 weeks, and ≥12 weeks
- >0 to <4 weeks, ≥4 weeks to <8 weeks, ≥8 to <12 weeks, and ≥12 weeks

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

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

These are described in detail in compound-level safety standards.

6.13.2. Adverse Events

Adverse events are recorded in the eCRFs. The planned summaries are provided in Table JAIV.6.7 and are described more fully in compound-level safety standards and in the adverse event-related PhUSE white paper (Analysis and Displays Associated with Adverse Events: Focus on Adverse Events in Phase 2-4 Clinical Trials and Integrated Summary Document [PhUSE 2017]). The analysis population and period are defined in Table JAIV.6.7.

Table JAIV.6.7. Summary Tables Related to Adverse Events

Analysis	Analysis Set
An overview table, with the number and percentage of patients with 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	T, T+FU
The percentages of patients with TEAEs will be summarized by treatment using MedDRA Preferred Term nested within System Organ Class.	T
The percentages of patients with TEAEs will be summarized by treatment using MedDRA Preferred Term (without regard to System Organ Class).	T
The percentages of patients with TEAEs will be summarized by treatment using MedDRA Preferred Term for the common TEAEs (occurred in >=2% [before rounding] of treated patients).	T
The percentages of patients with TEAEs by maximum severity will be summarized by treatment using MedDRA Preferred Term for the common TEAEs. Only counts and percentages will be included for the TEAEs by maximum severity.	T
A listing of all deaths from randomization to end of study participation in each study will be provided. Additional deaths that are reported outside of the study period will be obtained from the Lilly Safety System (LSS).	Enrolled
The number and percentage of patients who experienced a SAE (including deaths and SAEs temporally associated or preceding deaths) during the treatment period will be summarized by treatment using MedDRA Preferred Term nested within System Organ Class.	T
A listing of SAEs will be provided.	Enrolled
The number and percentage of patients who permanently discontinued from study drug due to an adverse event (including adverse events that led to death) during the treatment period will be summarized by treatment using MedDRA Preferred Term nested within System Organ Class.	T
The number and percentage of patients who temporarily interrupted study drug due to an adverse event during the treatment period will be summarized by treatment using MedDRA Preferred Term nested within System Organ Class.	T

Abbreviations: AE = adverse event; PT = preferred term; SAE = serious adverse event; T = treatment; T+FU = treatment plus follow-up; TEAE = treatment-emergent adverse event.

6.13.3. Deaths, Other Serious Adverse Events, and Other Notable Adverse Events

The planned table and listing for SAEs are included in the previous section. A listing of deaths, regardless of when they occurred during the study, will also be provided.

6.13.4. Adverse Events of Special Interest

6.13.4.1. Abnormal Hepatic Tests

Hepatic labs include alanine aminotransferase (ALT) and aspartate transaminase (AST), total bilirubin (TBL) and serum alkaline phosphatase (ALP). When criteria are met for hepatic evaluations, investigators will complete a follow-up hepatic safety eCRF. The planned

summaries are provided in Table JAIV.6.8 and are described more fully in compound-level safety standards.

Table JAIV.6.8. Summary Tables and Figures Related to Hepatic Safety

Analysis	Analysis Set
ALT and AST: The percentages of patients with a measurement greater than or equal to 3 times (3X), 5 times (5X), and 10 times (10X) the central lab upper limit of normal (ULN) during the treatment period for all patients with a post-baseline value and for subsets based on various levels of baseline value.	T
TBL: The percentages of patients with a measurement greater than or equal to 2 times (2X) the central lab 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 value.	T
ALP: The percentages of patients with a measurement greater than or equal to 2 times (2X) the baseline value and a measurement greater than or equal to 3 times (3X) the baseline value during the treatment period will be summarized for all patients with a post-baseline value and for subsets based on various levels of baseline value.	T
eDISH plot of maximum post-baseline ALT divided by ULN vs. maximum post-baseline total bilirubin divided by ULN.	T+FU
Patient profiles including demographics, disposition, information collected on the hepatic-safety CRF (where applicable) and a display of study drug exposure, adverse events, medications, blood pressure, heart rate, and the liver-related measurements over time will be provided for patients with information collected on the hepatic-safety CRF and any additional patients meeting ALT or AST measurement greater than or equal to 5X ULN (on a single measurement), or ALP measurement greater than or equal to 3X baseline ALP (on a single measurement), or ALP measurement greater than or equal to 2X baseline ALP on two or more consecutive measurements.	T+FU

Abbreviations: ALP = alkaline phosphatase; ALT = alanine aminotransferase; AST = aspartate transaminase; T = treatment; T+FU = treatment plus follow-up; TBL = total bilirubin.

6.13.4.2. Hematologic Changes

Hematologic changes will be assessed through analysis of hemoglobin, white blood cell count, absolute neutrophil count, lymphocyte count, and platelet count. The planned summaries are provided in Table JAIV.6.9, and are described more fully in compound-level safety standards.

Table JAIV.6.9. Summary Tables Related to Hematologic Changes

Analysis	Analysis Set
Shift tables showing the number and percentage of patients based on baseline to maximum during the treatment period will be created, with baseline depicted by the most extreme CTCAE grade during the baseline period. With each shift table, a summary displaying the number and percentage of patients who decreased, increased, or stayed the same in CTCAE grade category will be presented.	T
The percentages of patients with treatment-emergent laboratory abnormalities at any time during the treatment period will be summarized, based on any increase in postbaseline CTCAE grade, increase to CTCAE Grade 1 or above, Grade 2 or above, Grade 3 or above, and Grade 4.	T
The percentages of patients with treatment-emergent thrombocytosis will be summarized, defined as an increase in platelet count from a maximum baseline value ≤600 billion/L to any postbaseline value >600 billion/L.	T
Listing of patients with treatment-emergent thrombocytosis may be provided	T+FU

Abbreviations: CTCAE = Common Terminology Criteria for Adverse Events; T = treatment; T+FU = treatment plus follow-up.

6.13.4.3. Lipid Effects

Lipid effects will be assessed through analysis of elevated total cholesterol, elevated LDL cholesterol, decreased and increased HDL cholesterol, and elevated triglycerides. The planned summaries are provided in Table JAIV.6.10 and are described more fully in compound-level safety standards.

Table JAIV.6.10. Summary Tables Related to Lipid Effects

Analysis	Analysis Set
Shift tables showing the number and percentage of patients based on baseline to maximum during the treatment period will be created, with baseline depicted by the most extreme NCEP-based level during the baseline period. With each shift table, a summary displaying the number and percentage of patients who decreased, increased, or stayed the same in NCEP-based level will be presented.	T
The percentages of patients with treatment-emergent shifts at any time during the treatment period will be summarized, based on increases to various levels of NCEP-based categories.	T

Abbreviation: NCEP = National Cholesterol Education Programme.

6.13.4.4. Renal Function Effects

Effects on renal function will be assessed through analysis of elevated creatinine. Renal events will be identified using terms from the acute renal failure SMQ. The planned summaries are provided in Table JAIV.6.11 and are described more fully in compound-level safety standards.

Table JAIV.6.11. Summary Tables Related to Effects on Renal Function

Analysis	Analysis Set
Shift tables showing the number and percentage of patients based on baseline to maximum during the treatment period will be created, with baseline depicted by the most extreme CTCAE grade during the baseline period. With each shift table, a summary displaying the number and percentage of patients who decreased, increased, or stayed the same in CTCAE grade category will be presented.	T
The percentages of patients with treatment-emergent laboratory abnormalities at any time during the treatment period will be summarized, based on any increase in postbaseline CTCAE grade, increase to CTCAE Grade 1 or above, Grade 2 or above, Grade 3 or above, and Grade 4.	T

Abbreviations: CTCAE = Common Terminology Criteria for Adverse Events; T = treatment; T+FU = treatment plus follow-up.

6.13.4.5. Evaluations in Creatine Phosphokinase (CPK)

The planned summaries are provided in Table JAIV.6.12 and are described more fully in compound-level safety standards.

Table JAIV.6.12. Summary Tables Related to Effects on Renal Function

Analysis	Analysis Set
Shift tables showing the number and percentage of patients based on baseline to maximum during the treatment period will be created, with baseline depicted by the most extreme CTCAE grade during the baseline period. With each shift table, a summary displaying the number and percentage of patients who decreased, increased, or stayed the same in CTCAE grade category will be presented.	T
The percentages of patients with treatment-emergent laboratory abnormalities at any time during the treatment period will be summarized, based on any increase in postbaseline CTCAE grade, increase to CTCAE Grade 1 or above, Grade 2 or above, Grade 3 or above, and Grade 4.	T

Abbreviations: CTCAE = Common Terminology Criteria for Adverse Events; T = treatment; T+FU = treatment plus follow-up.

6.13.4.6. Infections

Infections will be defined using all the PTs from the MedDRA Infections and Infestations SOC. The MedDRA terms used to identify infections considered to be opportunistic infections (OIs) are based on Winthrop et al. (2015) and are listed in the compound-level safety standards. The list contains narrow (more specific) and broad (less specific) PTs.

The planned summaries are provided in Table JAIV.6.13 and are described more fully in compound-level safety standards.

Table JAIV.6.13. Summary Tables Related to Infections

Analysis	Analysis Set
The number and percentage of patients with treatment-emergent infections, serious infections, and infections resulting in permanent study drug discontinuation using MedDRA PTs	T
The number and percentage of patients with TEAEs of infections by maximum severity using MedDRA PTs	T
Listing of patients experiencing TEAE infections will be provided. The listing will include patient demographics, treatment group, treatment start and stop dates, infectious PT event, event start and stop dates, total leukocytes, total lymphocytes, absolute neutrophils, event seriousness, and event outcome	T+FU
Listings and summary of OIs based on MedDRA PTs	T
A summary table of herpes zoster will be provided, including event maximum severity, seriousness, whether resulting in temporary study drug interruption, whether resulting in study drug discontinuation, whether treated with antiviral medication, and event outcome. The incidence rate adjusted for observation time will also be provided.	T
A listing of patients with detectable HBV DNA	T+FU
Hepatitis B virus DNA status (not detectable, detectable but not quantifiable [ie, < lower limit of detection (LLOD)], quantifiable [ie, ≥LLOD]) stratified by applicable baseline HBV serology status	T

Abbreviations: OIs = opportunistic infections; PTs = preferred terms; T = treatment; T+FU = treatment plus follow-up; TEAE = treatment-emergent adverse events.

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

Major adverse cardiovascular events (MACE) and other cardiovascular events will be adjudicated by an independent, external adjudication committee. All confirmed events after adjudication will be used for the analysis.

The planned summaries are provided in Table JAIV.6.14 and are described more fully in compound-level safety standards.

Table JAIV.6.14. Summary Tables Related to MACE and Other Cardiovascular Events

Analysis	Analysis Set
The number and percentage of patients with MACE, other cardiovascular events, non-cardiovascular death, and all-cause death, as adjudicated, will be summarized by treatment group based on the categories and subcategories as defined in compound-level safety standards.	T
A listing of the events sent for cardiovascular 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.	T+FU

Abbreviations: MACE = Major Adverse Cardiovascular Events; PT = preferred term ; T = treatment; T+FU = treatment plus follow-up.

6.13.4.8. Venous and Pulmonary Artery Thromboembolic (VTE) Events

Venous thromboembolism (VTE) events will be adjudicated by an independent, external adjudication committee. Venous and pulmonary artery thromboembolic events will be classified as deep vein thrombosis (DVT), pulmonary embolism (PE), or other peripheral venous thrombosis. All confirmed events after adjudication will be used for the analysis.

The planned summaries are provided in Table JAIV.6.15 and are described more fully in compound-level safety standards.

Table JAIV.6.15. Summary Tables Related to VTE Events

Analysis	Analysis Set
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.	T
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.	T+FU

Abbreviations: DVT = deep vein thrombosis; PE = pulmonary embolism; PT = preferred term; T = treatment; T+FU = treatment plus follow-up; VTE = venous thromboembolism.

6.13.4.9. Arterial Thromboembolic (ATE) Events

Standardized MedDRA Queries (SMQs) or predefined search methods will be used to identify the candidate cases. These candidate cases will be confirmed by medical review assessment.

The planned summaries are provided in Table JAIV.6.16 and are described more fully in compound-level safety standards.

Table JAIV.6.16. Summary Tables Related to ATE Events

Analysis	Analysis Set
The number and percentage of patients with treatment emergent ATE will be summarized by treatment group.	T
Patient profile reports and LSS summaries will be provided for all treatment-emergent serious and non-serious ATE.	T

Abbreviations: ATE = Arterial thromboembolic event; LSS = Lilly Safety System; T = treatment; T+FU = treatment plus follow-up.

6.13.4.10. Malignancies

Malignancies will be identified using terms from the malignant tumors SMQ. Malignancies excluding nonmelanoma skin cancers (NMSC) and NMSC will be reported separately. All the cases identified by the malignant tumors SMQ will be assessed through medical review to determine *confirmed* NMSC cases.

The planned summaries are provided in Table JAIV.6.17 and are described more fully in compound-level safety standards.

Table JAIV.6.17. Summary Tables Related to Malignancies

Analysis	Analysis Set
The number and percentage of patients with treatment-emergent malignancies excluding NMSC and NMSC will be summarized by treatment group.	T
Listing of all malignancy cases, with an NMSC flag.	T+FU

Abbreviation: NMSC = nonmelanoma skin cancers; T = treatment; T+FU = treatment plus follow-up.

6.13.4.11. Allergic Reactions/Hypersensitivities

A search for relevant events related to allergic reaction and hypersensitivity will be performed using the following SMQs:

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

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.

The summaries described in Table JAIV.6.18 will be created if there are sufficient numbers of events to warrant further examination beyond the listing specified above.

Table JAIV.6.18. Summary Tables Related to Allergic Reactions/Hypersensitivities

Analysis	Analysis Set
The number and percentage of patients with TEAEs will be summarized by treatment using MedDRA Preferred Term for any narrow or algorithmic term from any one of the 3 SMQs (each SMQ and SMQs combined)	T
The number and percentages of patients with TEAEs will be summarized by treatment using MedDRA Preferred Term for any narrow scope term (each SMQ separately)	T
The number and percentages of patients with TEAEs will be summarized by treatment using MedDRA Preferred Term for any broad term (each SMQ separately)	T

Abbreviations: TEAE = treatment-emergent adverse event; SMQ = standardized MedDRA queries; T = treatment.

6.13.4.12. Gastrointestinal Perforations

Potential gastrointestinal perforations will be identified using terms from the GI perforations SMQ. Potential GI perforations identified by the 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 GI perforation. All confirmed events after medical review will be used for later analysis.

6.13.4.13. Depression and Suicide

During the study, suicidal ideation and behavior, and depression, will be assessed prospectively by the investigator via signs and symptoms and through the use of the Quick Inventory of Depressive Symptomatology Self Report (QIDS-SR16). The QIDS-SR16 total scores will also be categorized into severity classes (none, mild, moderate, severe, very severe).

The planned summaries are provided in Table JAIV.6.19 and are described more fully in compound-level safety standards.

Table JAIV.6.19. Summaries Related to Depression and Suicide

Analysis	Analysis Set
Treatment differences in mean change QIDS-SR_{16} total score will be analyzed.	T
Frequency counts and percentages will be used to summarize the QIDS-SR_{16} total score severity categories at each visit by treatment group.	T
Shift tables showing the number and percentage of patients based on baseline to maximum during the treatment period will be created, with baseline depicted by the most extreme QIDS-SR16 category during the baseline period. With each shift table, a summary displaying the number and percentage of patients who decreased, increased, or stayed the same in QIDS-SR16 category will be presented.	T
Shift tables will be created for the QIDS-SR16 suicidal ideation item responses.	T
The percentages of patients with treatment-emergent shifts at any time during the treatment period will be summarized, based on any increase to mild and above, moderate or above, severe or above, very severe.	T

Abbreviation: QIDS-SR16 = Quick Inventory of Depressive Symptomatology Self Report; T = treatment.

6.13.5. Clinical Laboratory Evaluation

The planned summaries for clinical laboratory evaluations are provided in Table JAIV.6.20 and are described more fully in compound-level safety standards and in the laboratory-related PhUSE white papers [Analyses and Displays Associated with Measures of Central Tendency – Focus on Vital Sign, Electrocardiogram, and Laboratory Analyte Measurements in Phase 2-4 Clinical Trials and Integrated Submission Documents (PhUSE 2013) and Analyses and Displays Associated with Outliers or Shifts from Normal to Abnormal: Focus on Vital Signs, Electrocardiogram, and Laboratory Analyte Measurements in Phase 2-4 Clinical Trials and Integrated Summary Documents (PhUSE 2015)].

Table JAIV.6.20. Summary Tables Related to Clinical Laboratory Evaluations

Analysis	Analysis Set
Box plots for observed values by visit and change from baseline values by visit and at last observation.	T excluding visit 801
Tables with percentages of patients who shift from normal/high to low (i.e., treatment-emergent low) and percentages of patients who shift from normal/low to high (i.e., treatment-emergent high)	T
Listing of abnormal findings for laboratory analyte measurements, including qualitative measures	T+FU

Abbreviations: T = treatment; T+FU = treatment plus follow-up.

6.13.6. Vital Signs and Other Physical Findings

The planned summaries for vital signs and physical characteristics (systolic BP, diastolic BP, pulse, weight, BMI and waist circumference) are provided in Table JAIV.6.21 and are described more fully in compound-level safety standards and in the vitals-related PhUSE white papers [Analyses and Displays Associated with Measures of Central Tendency – Focus on Vital Sign, Electrocardiogram, and Laboratory Analyte Measurements in Phase 2-4 Clinical Trials and Integrated Submission Documents (PhUSE 2013) and Analyses and Displays Associated with Outliers or Shifts from Normal to Abnormal: Focus on Vital Signs, Electrocardiogram, and Laboratory Analyte Measurements in Phase 2-4 Clinical Trials and Integrated Summary Documents (PhUSE 2015)].

Table JAIV.6.21. Summary Tables Related to Vital Signs

Analysis	Analysis Set
Box plots for observed values by visit and change from baseline values by visit and at last observation.	T excluding visit 801
Tables with percentages of patients who shift from normal/high to low (i.e., treatment-emergent low) and percentages of patients who shift from normal/low to high (i.e., treatment-emergent high). The limits are defined in the compound-level safety standards and are based on literature.	T

Abbreviation: T = treatment.

6.14. Subgroup Analyses

Subgroup analyses comparing each dose of baricitinib to placebo will be performed on the mITT population at Week 12, with data up to rescue for the following:

- Change from baseline in alkaline phosphatase (ALP) at Week 12.

The following subgroups (but not necessarily limited to only these) will be evaluated:

- [redacted]

Descriptive statistics will be provided for each treatment and stratum of a subgroup as outlined, regardless of sample size. The analysis will be performed using MMRM. The model will include treatment, subgroup, and treatment-by-subgroup interaction as explanatory variables. The treatment-by-subgroup interaction comparing treatment groups will be tested at the 0.1 significance level. The p-value from the MMRM will be reported for the interaction test. Response counts and percentages will be summarized by treatment for each subgroup category. Type III tests will be used for all the statistical comparisons; the 90% confidence interval (CI) will also be reported. The LS mean difference, standard error, p-value, and CI will also be reported. Treatment group comparisons at specific study visits will be tested using the t-test obtained from the MMRM results. Additionally, descriptive statistics will also be provided for each treatment and cohort (Cohort A and Cohort B) as a subgroup.

6.15. 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.

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

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.

6.16. Interim Analyses and Data Monitoring

An interim analysis will be conducted under the auspices of a DMC according to the specifications set forth in the protocol.

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. The DMC will include at least 3 individuals including a chairperson independent of Lilly (external to Lilly) and independent of any third party organization (TPO) contracted by Lilly to conduct the study. The DMC will include at least 2 physicians (at least one specialist with expertise in hepatology) with experience as a DMC member, as well as a biostatistician with DMC and clinical trial experience.

The DMC will be authorized to review unblinded results of analyses by treatment group prior to database lock, including study discontinuation data, AEs including SAEs, clinical laboratory data, vital sign data, etc. The DMC may recommend to continue the study as designed; temporarily suspend enrollment, pending resolution of a specified issue; or discontinue the entire study. While the DMC may request to review efficacy data to investigate the benefit/risk relationship in the context of safety observations for ongoing patients in the study, no information regarding efficacy will be communicated. Moreover, the study will not be stopped for positive efficacy results nor will it be stopped for futility. Hence, no alpha is spent. Details of the DMC, including its operating characteristics, are documented in the Baricitinib Data Monitoring Committee Charter for Protocol I4V-MC-JAIV, and further details are given in the Interim Analysis Plan in Section 6.16.1.

Besides DMC members, a limited number of pre-identified individuals may gain access to the limited unblinded data, as specified in the unblinding plan, prior to the interim or final database lock, to initiate the final population PK/PD model development processes. Information that may unblind the study during the analyses will not be reported to study sites or blinded study team until the study has been unblinded.

Unblinding details are given in a separate blinding and unblinding plan document.

6.16.1. Interim Analysis Plan

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

- patient disposition, demographics, and baseline characteristics
- exposure
- timing of and reasons for study discontinuation and permanent study drug discontinuation
- adverse events (AEs), to include the following:
    - treatment-emergent AEs
    - serious AEs, including deaths
    - selected special safety topics
- clinical laboratory results
- vital signs

All listings will include patient ID and treatment group. Summaries will include TEAEs, SAEs, special topics AEs, and treatment-emergent high and low laboratory and vital signs in terms of counts, percentages and incidence rates where applicable. For continuous analyses, box plots of laboratory analytes will be provided by time point and summaries will include descriptive statistics such as mean, SD, minimum, 1st quartile, median, 3rd quartile, and maximum.

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 study. If efficacy data are requested, change from baseline in ALP will be provided. Further details are given in the DMC charter.

6.17. Planned Exploratory Analyses

The planned exploratory analyses are described in Sections 6.10 and 6.11. 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.

6.18. Annual Report Analyses

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

6.19. 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 the following:

Summary of adverse events, provided as a dataset which will be converted to an XML file. Both Serious Adverse Events and "Other" Adverse Events are summarized: by treatment group, by MedDRA preferred term.

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

7. Unblinding Plan

Refer to a separate blinding and unblinding plan document for details.

8. References

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

[CTCAE] Common Terminology Criteria for Adverse Events, Version 4.03, DHHS NIH NCI, NIH Publication No. 09-5410, June 14, 2010. Available at: https://evs.nci.nih.gov/ftp1/CTCAE/CTCAE_4.03_2010-06-14_QuickReference_5x7.pdf. Accessed 25 July 2017.

FACIT Fatigue Scale (Version 4). Available at: http://www.ser.es/wp-content/uploads/2015/03/FACIT-F_INDICE.pdf. Accessed 08 May 2019.

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

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

Jacoby A, Rannard A, Buck D, Bhala N, Newton JL, James OFW, Jones DEJ. Development, validation, and evaluation of the PBC-40, a disease specific health related quality of life measure for primary biliary cirrhosis. *Gut*. 2005;54(11):1622-1629.

PhUSE Computational Science Deliverables Catalog. Analysis and Displays Associated with Adverse Events: Focus on Adverse Events in Phase 2-4 Clinical Trials and Integrated Summary Document. 2017. https://www.phuse.eu/documents//working-groups/deliverables/adverse-events-white-paper-version-10-03-feb-17-11796-19835.pdf

PhUSE Computational Science Deliverables Catalog. Analyses and Displays Associated with Measures of Central Tendency – Focus on Vital Sign, Electrocardiogram, and Laboratory Analyte Measurements in Phase 2-4 Clinical Trials and Integrated Submissin Documents. (2013).http://www.phusewiki.org/wiki/images/4/48/CSS_WhitePaper_CentralTendency_v1.0.pdf.

PhUSE Computational Science Deliverables Catalog. Analyses and Displays Associated with Outliers or Shifts from Normal to Abnormal: Focus on Vital Signs, Electrocardiogram, and Laboratory Analyte Measurements in Phase 2-4 Clinical Trials and Integrated Summary Documents. (2015). https://www.phuse.eu/documents//working-groups/deliverables/analyses-displays-associated-with-outliers-or-shifts-from-normal-to-abnormal-version-10-10-sept-15-11815.pdf.

Program Safety Analysis Plan: Baricitinib (LY3009104) Version 6. Report on file, Eli Lilly and Company.

Protocol I4V-MC-JAIV(b). A Randomized, Double-Blind, Placebo-Controlled, Proof-of-Concept Study Evaluating the Efficacy and Safety of Baricitinib (LY3009104) in

Patients with Primary Biliary Cholangitis Who Have an Inadequate Response or Are Intolerant to UDCA. Report on file, Eli Lilly and Company.

Webster K, Cella D, Yost K. The Functional Assessment of Chronic Illness Therapy (FACIT) measurement system: properties, applications, and interpretation. *Health Qual Life Outcomes*. 2003;1:79.