Statistical Analysis Plan I4V-CR-JAGS

Randomized, Double-Blind, Placebo-Controlled, Phase 3 Study Evaluating the Efficacy and Safety of Baricitinib in Patients with Moderately to Severely Active Rheumatoid Arthritis Who Have Had an Inadequate Response to Methotrexate Therapy

NCT02265705

Approval Date: 12-Jan-2017

1. Statistical Analysis Plan: I4V-CR-JAGS: A Randomized, Double-Blind, Placebo-Controlled, Phase 3 Study Evaluating the Efficacy and Safety of Baricitinib in Patients with Moderately to Severely Active Rheumatoid Arthritis Who Have Had an Inadequate Response to Methotrexate Therapy

Confidential Information

Baricitinib (LY3009104) Rheumatoid Arthritis

Study I4V-CR-JAGS is a 52-week, Phase 3, multicenter, randomized, double-blind, placebo-controlled, parallel-group study comparing the efficacy of baricitinib versus placebo in patients with moderately to severely active rheumatoid arthritis who have had an insufficient response (IR) to methotrexate (MTX) and who have never been treated with a biologic disease-modifying antirheumatic drug (that is, MTX-IR patients).

Eli Lilly and Company
Indianapolis, Indiana USA 46285
Protocol I4V-CR-JAGS
Phase 3

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

2. Table of Contents

Section **Page**

Table of Contents

Table

Table of Contents

Figure **Page**

Table of Contents

Appendix **Page**

3. Revision History

This statistical analysis plan (SAP) Version 1 was approved prior to unblinding.

4. Study Objectives

The following objectives are replicated from Amendment (c) of Study I4V-CR-JAGS (JAGS) protocol, dated 07 Dec 2016.

4.1. Primary Objective

The primary objective of the study is to determine whether baricitinib is superior to placebo in the treatment of patients with moderately to severely active rheumatoid arthritis (RA) despite methotrexate treatment (that is, MTX-IR), as assessed by the proportion of patients achieving a 20% improvement in the American College of Rheumatology (ACR) criteria (ACR20) at Week 12.

4.2. Secondary Objectives

The secondary objectives of the study are to evaluate the efficacy of baricitinib versus placebo as assessed by:

- change from baseline to Week 12 in Health Assessment Questionnaire-Disability Index (HAQ-DI) score
- change from baseline to Week 12 in Disease Activity Score modified to include the 28 diarthrodial joint count (DAS28)-high-sensitivity C-reactive protein (hsCRP)
- proportion of patients achieving a Simplified Disease Activity Index (SDAI) score ≤3.3 at Week 12
- mean duration of morning joint stiffness in the 7 days prior to Week 12 as collected in diaries
- mean severity of morning joint stiffness in the 7 days prior to Week 12 as collected in diaries
- mean Worst Tiredness numeric rating scale (NRS) in the 7 days prior to Week 12 as collected in diaries
- mean Worst Joint Pain NRS in the 7 days prior to Week 12 as collected in diaries.

4.3. Exploratory Objectives

The exploratory objectives in the study include efficacy evaluation of baricitinib as assessed by:

- change from baseline at Weeks 16, 24, and 52 in structural joint damage as measured by modified Total Sharp Score (mTSS; van der Heijde 2000)
- proportion of patients with mTSS change ≤0 from baseline at Weeks 16, 24, and 52
- change from baseline to Weeks 16, 24, and 52 in joint space narrowing and bone erosion scores
- proportion of patients achieving DAS28-hsCRP ≤3.2 and DAS28-hsCRP <2.6 at Weeks 12, 24, and 52
- proportion of patients achieving HAQ-DI improvement ≥0.22 and ≥0.3 at Weeks 12, 24, and 52
- proportion of patients achieving ACR20 at Week 24 and Week 52
- proportion of patients achieving 50% improvement in ACR criteria (ACR50) at Weeks 12, 24, and 52
- proportion of patients achieving 70% improvement in ACR criteria (ACR70) at Weeks 12, 24, and 52

- percentage change from baseline to Weeks 12, 24, and 52 in individual components of the ACR Core Set including tender joint count based on 68 joint counts (TJC68), swollen joint count based on 66 joint counts (SJC66), patient's assessment of pain, Patient's Global Assessment of Disease Activity, Physician's Global Assessment of Disease Activity, HAQ-DI, and hsCRP
- change from baseline through Week 52 in DAS28-hsCRP
- change from baseline through Week 52 in DAS28-erythrocyte sedimentation rate (ESR)
- proportion of patients achieving DAS28-ESR ≤3.2 and DAS28-ESR <2.6 at Weeks 12, 24, and 52
- change from baseline to Week 24 and Week 52 in HAQ-DI score
- change from baseline to Weeks 12, 24, and 52 in Clinical Disease Activity Index (CDAI) score
- change from baseline to Weeks 12, 24, and 52 in SDAI score
- proportion of patients achieving a CDAI score ≤2.8 at Weeks 12, 24, and 52
- proportion of patients achieving an SDAI score ≤3.3 at Weeks 24 and 52
- proportion of patients achieving ACR/European League Against Rheumatism (EULAR) remission according to the Boolean-based definition at Weeks 12, 24, and 52
- proportion of patients achieving moderate and good EULAR responses based on the DAS28 at Weeks 12, 24, and 52
- evaluation of hybrid ACR (bounded) response at Weeks 12, 24, and 52
- change from baseline to Week 12, Week 24, and Week 52 in Functional Assessment of Chronic Illness Therapy-Fatigue (FACIT-F) scale scores
- change from baseline to Weeks 12, 24, and 52 in European Quality of Life-5 Dimensions-5 Level (EQ-5D-5L) scores
- change from baseline to Weeks 12, 24, and 52 in Work Productivity and Activity Impairment-Rheumatoid Arthritis (WPAI-RA) scores
- evaluation of healthcare resource utilization through Week 52
- change from baseline to Weeks 12, 24, and 52 in duration of morning joint stiffness assessed at the visit using an electronic patient-reported outcomes (ePRO) tablet
- change from baseline to Weeks 12, 24, and 52 in Worst Tiredness NRS assessed at the visit using an ePRO tablet
- change from baseline to Weeks 12, 24, and 52 in Worst Joint Pain NRS assessed at the visit using an ePRO tablet.

5. Study Design

5.1. Summary of Study Design

Study JAGS is a 52-week, Phase 3, multicenter, randomized, double-blind, placebo-controlled, parallel-group study comparing the efficacy of baricitinib versus placebo in patients with moderately to severely active RA who have had an insufficient response to MTX and who have never been treated with a biologic disease-modifying antirheumatic drug (bDMARD) (that is, MTX-IR patients).

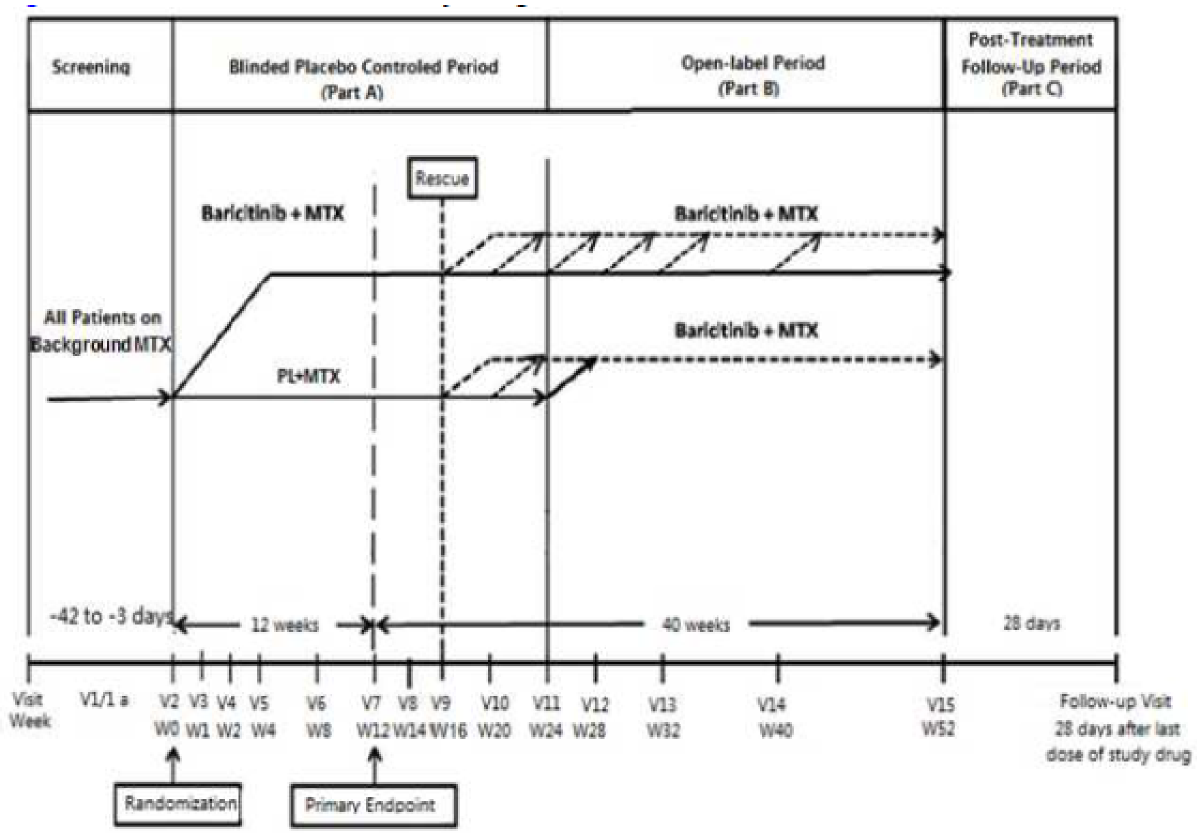
Baricitinib will be administered as a 4-mg tablet once daily (QD). The dose for patients with renal impairment, defined as estimated glomerular filtration rate (eGFR) <60 mL/min/1.73 m^2, will be 2 mg baricitinib QD. Patients not assigned to the baricitinib treatment arm will receive either a 4-mg or 2-mg matching placebo tablet.

Planned enrollment is approximately 288 randomized patients: 144 to receive baricitinib, and 144 to receive placebo. After the Week 52 study visit, eligible patients may proceed to a separate extension study (Study I4V-MC-JADY [JADY]) lasting for up to 49 months or to the post-treatment follow-up period of this study (Part C).

Study JAGS will consist of 4 parts:

- Screening: Screening period lasting from 3 to 42 days prior to Visit 2 (Week 0)
- Part A: double-blind, placebo-controlled period from Week 0 through Week 24
- Part B: open-label period from the end of Week 24 through Week 52
- Part C: post-treatment follow-up period (28 days)

Figure 5.1 illustrates the study design.



Abbreviations: MTX = methotrexate, V = visit, W = week.

Note: Diagonal dashed arrows indicate an option for rescue therapy. The diagonal solid arrow indicates a mandatory change to baricitinib treatment (at Week 24 for placebo-treated patients).

Figure 5.1. Illustration of study design for Clinical Protocol I4V-CR-JAGS.

5.1.1. *Screening Period*

This is a screening period of not less than 3 days and not more than 42 days. In exceptional circumstances, the screening window can be extended after consultation with the Sponsor. Patients will be screened for study eligibility at Visit 1 and will be randomized at Visit 2 to continue in Part A.

5.1.2. *Parts A and B: 24-Week Double-Blind Placebo-Controlled (Part A) Period and 28-Week Open-Label (Part B) Period*

Patients on a background of MTX will be randomized at a 1:1 ratio to 1 of 2 treatment arms (baricitinib or placebo):

- Patients randomized to baricitinib will be administered a 4-mg tablet QD (or 2 mg baricitinib tablet QD if eGFR <60 mL/min/1.73 m^2) starting at Week 0 (Visit 2) through Week 52 (Visit 15).
- Patients randomized to placebo will be administered a placebo tablet QD starting at Week 0 (Visit 2) through Week 24 (the morning of Visit 11). At Week 24, patients will be administered baricitinib 4 mg QD (or 2 mg baricitinib QD if eGFR <60 mL/min/1.73 m^2)

through Week 52 (Visit 15). Patients will be eligible for rescue treatment beginning at Week 16 (Visit 9). Details of the rescue are listed in Section 5.3.3.

Patients will remain on background MTX, nonsteroidal anti-inflammatory drugs (NSAIDs), analgesics, and/or corticosteroids (see Protocol Section 9.8 for guidelines on adjustments to concomitant therapy). Patients will take their assigned doses of study drugs as directed. The primary efficacy endpoint will be at Week 12, and the final visit during Part A will be at Week 24. The final visit during Part B will be at Week 52 (Visit 15). The structural joint damage endpoints will be obtained at Weeks 16, 24, and 52. All patients who complete the Week 52 study visit will be eligible for inclusion in the extension Study JADY.

5.1.3. *Part C: Post-Treatment Follow-Up Period*

Patients who are not enrolled in the extension Study JADY will have a Follow-Up Visit (Visit 801) approximately 28 days after the last dose of study drug.

Patients who discontinue from the study prior to Week 52 for any reason will have an Early Termination Visit performed, including x-rays. The Early Termination Visit should be performed as soon as logistically possible because this visit includes assessment of safety. Infrequently, patient and investigator availability may be such that the Early Termination Visit and the Follow-up Visit may occur on or about the same date. In this instance, the visits may be combined and should occur approximately 28 days after the last dose of study drug. All activities required for the Early Termination Visit should be completed according to the schedule of assessments (Protocol Attachment 1).

5.1.4. *Study Extensions*

Patients who complete this study through Visit 15 will be eligible to participate in Study JADY, if enrollment criteria for Study JADY are met. These patients will not participate in Part C.

5.2. Determination of Sample Size

Approximately 288 patients (144 in baricitinib, 144 in placebo) will provide:

- 99% power to detect a difference between the baricitinib and placebo treatment groups in ACR20 response rate (60% versus 35%) at Week 12 based on a 2-sided chi-square test at a significance level of 0.05.

The sample size and power estimates were obtained using nQuery® Advisor 7.0.

5.3. Method of Assignment to Treatment

Patients who meet all criteria for enrollment will be randomized in a 1:1 ratio (baricitinib:placebo) to double-blind treatment at Week 0 (Visit 2). Randomization will be stratified by country and joint erosion status (1-2 joint erosions plus seropositivity versus at least 3 joint erosions). Assignment to treatment groups will be determined by a computer-generated random sequence using the interactive web-response system (IWRS).

The IWRS will be used to assign packages of baricitinib/placebo containing double-blind study drug to each patient. Each package of clinical trial material will supply sufficient medication for the number of weeks between visits. Site personnel will confirm that they have located the correct package by entering a confirmation number found on the packages of study drug into the IWRS before dispensing the package to the patient.

5.3.1. *Treatment Administered*

This study involves a comparison of baricitinib 4 mg administered orally QD and placebo administered orally QD. Patients will continue to take background MTX therapy during the course of the study. Table 5-1 shows the treatment regimens.

Table 5-1 Treatment Regimens

Treatment Group	Treatments Administered during Part A (Patients Not Rescued)	Treatments Administered during Part B
Baricitinib treatment	Baricitinib 4 mg[a] oral once-daily tablet	Baricitinib 4 mg[a] oral once-daily tablet
Placebo comparator	Baricitinib placebo oral once-daily tablet	Baricitinib 4 mg[a] oral once-daily tablet

[a] The dose for patients with renal impairment, defined as estimated glomerular filtration rate <60 mL/min/1.73 m^2, will be 2 mg once daily.

The investigator or his/her designee will be responsible for explaining the correct use of study drug(s) to the patient, verifying that instructions are followed properly, maintaining accurate records of study drug dispensing and collection, and returning all unused medication to Eli Lilly and Company (Lilly) or its designee at the end of the study.

Patients will be instructed to contact the investigator as soon as possible if they have a complaint or problem with the study drug so that the situation can be assessed.

5.3.2. *Selection and Timing of Doses*

Each patient will receive blinded study drug (baricitinib 4 mg or baricitinib placebo tablets orally QD) beginning on the day of Visit 2 (Week 0). Baricitinib 4-mg oral dosing will continue daily through Visit 15 (Week 52). Baricitinib placebo oral dosing will continue daily through Visit 11 (Week 24); at Visit 11 (Week 24), all patients randomized to placebo will be reassigned to baricitinib.

Oral study drug should be taken at home at approximately the same time each day, as much as possible:

- At Visit 2 (Week 0), patients will take their study drug in the clinic.

5.3.3. *Rescue Therapy*

In consideration of the disease severity, all patients in Study JAGS will be offered rescue therapy starting at Week 16 if they are determined to be nonresponders. Nonresponse at Week 16 is defined as lack of improvement of at least 20% in both tender joint counts (TJCs) and swollen joint counts (SJCs) at both Week 14 and Week 16 compared to baseline. At Week 16, the IWRS will assign rescue treatment based on TJC and SJC values.

All patients who are non-responders (in both baricitinib and placebo arm) will receive rescue therapy at Week 16. Patients who are nonresponders and who were randomized to placebo will be rescued with baricitinib 4 mg. To maintain study integrity through study drug blinding, patients who were originally randomized to baricitinib will continue to receive baricitinib, with appropriate ongoing assessment of response and use of concomitant medication as needed.

After Week 16, rescue therapy will be offered to patients at the discretion of the investigator based on TJCs and SJCs. Once a patient has been rescued, new NSAIDs, corticosteroids, and/or analgesics may be added or doses of ongoing concomitant NSAIDs, corticosteroids, and/or analgesics may be increased at the discretion of the investigator.

A patient may only be rescued once. If a patient continues to meet nonresponse criteria for 4 weeks after rescue or at any time point thereafter, that patient should be discontinued from the study.

Patients initially randomized to placebo and not rescued before Week 24 will be reassigned to receive baricitinib 4 mg at Week 24 in order to limit exposure to inactive treatment.

Nonresponders who are rescued will receive an oral tablet daily for the remainder of the study. Rescue therapy will be administered through Week 52.

6. A Priori Statistical Methods

6.1. General Considerations

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

Statistical analysis of this study will be the responsibility of Lilly.

For continuous measures, summary statistics will include the sample size, means, standard deviations (SDs), medians, minimums, 1st quartiles (Q1), 3rd quartiles (Q3), and maximums. For categorical measures, summary statistics will include the sample size, frequency counts, and percentages.

All statistical tests of treatment effects will be performed at 2-sided significance levels of 0.05, unless otherwise stated.

All p-values will be rounded up to 3 decimal places. For example, any p-value strictly greater than 0.049 and less than or equal to 0.05 will be displayed as 0.050. This guarantees that on any printed statistical output, the unrounded p-value will always be less than or equal to the displayed p-value. A displayed p-value of 0.001 will always be understood to mean ≤0.001. Likewise, any p-value displayed as 1.000 will be understood to mean >0.999 and ≤1.

Lilly will use both an effectiveness estimand (that is, the relative effect attributable to the originally randomized treatments, baricitinib and placebo, at the primary time point in all treated patients, even if patients change treatment) and an efficacy estimand (that is, the relative effect attributable to the originally randomized treatments, baricitinib and placebo, at the primary time point in all treated patients if taken as directed) to evaluate efficacy measures at the primary time point of Week 12 for Study JAGS. All primary and secondary endpoints related to improvements in symptomatic features (that is, RA signs and symptoms, physical function, and patient-reported outcomes) are assessed prior to patients having the option to receive rescue therapy. Therefore, the only practical difference between the efficacy and effectiveness estimands for these endpoints concerns assumptions about patient response after permanent discontinuation of study drug or from the study prior to the analysis time point.

For outcome measures defined by response/nonresponse categories, nonresponder imputation (NRI) is utilized with both estimands. This approach considers study discontinuation to be a negative outcome that outweighs any potential benefit gained from hypothetical continued treatment or actual treatment and therefore fits in either the efficacy or effectiveness frameworks.

For continuous outcome measures that quantify a treatment effect, Lilly utilizes modified baseline observation carried forward (mBOCF). At the time of discontinuation, some patients may have positive outcomes (or even a range of effects) and discontinue for reasons unrelated to treatment; it seems reasonable to reflect those outcomes in the expected effects attributable to the originally randomized treatments. Other patients may discontinue because they were not seeing benefit, in which case their outcomes at the time of discontinuation should reflect their level of expected efficacy (or lack thereof) attributable to the originally randomized treatments. Finally, other patients may see positive outcomes (or even a range of effects) but be unable to tolerate or continue receiving the medicine because of adverse events (AEs). Because the latter is a negative outcome in spite of potentially positive efficacy data, these patients will be considered to have received no improvement in their symptomatic features; therefore, the patient's baseline observation is carried forward for analysis. These approaches are reasonable for use with both the efficacy and effectiveness estimands when the inference concerns the originally randomized treatments.

For Study JAGS, Lilly plans to use this described approach for evaluating the primary endpoints. Additional sensitivity analyses are planned to investigate the robustness of this approach.

Treatment comparisons of categorical efficacy variables (including proportions of patients achieving ACR20) will be made using a logistic regression analysis with country, baseline joint erosion status (1-2 joint erosions plus seropositivity versus at least 3 joint erosions), and treatment group in the model. In general, the p-value and 95% confidence interval (CI) for the odds ratio from the logistic regression model are used for primary statistical inference. When logistic regression sample size requirements (<5 responders in any category for any factor) are not met, the p-value from the Fisher exact test is produced instead of the odds ratio and 95% CI.

Treatment comparisons of continuous efficacy and health outcomes variables will be made using analysis of covariance (ANCOVA) with country, treatment group, baseline joint erosion status, and baseline value in the model. Type III sums of squares for the least-squares (LS) means will be used for the statistical comparison; the 95% CI will also be reported.

When a restricted maximum likelihood (REML)-based mixed model for repeated measures (MMRM) is used (as a sensitivity model for continuous secondary variables [see Section 6.11.2]), the model will include treatment, country, baseline joint erosion status (1-2 joint erosions plus seropositivity versus at least 3 joint erosions), visit, and treatment-by-visit-interaction as fixed categorical effects and baseline score and baseline score-by-visit-interaction as fixed continuous effects to estimate change from baseline across post-baseline visits. An unstructured covariance structure will be used to model the between- and within-patient errors. If this analysis fails to converge, other structures will be tested such as heterogeneous autoregressive [ARH(1)], heterogeneous compound symmetry (CSH), or heterogeneous Toeplitz (TOEPH). The variance-covariance structure that results in the smallest Akaike information criterion (AIC) will be used. The Kenward-Roger method will be used to estimate the denominator degrees of freedom. Type III sums of squares for the LS means will be used for the statistical comparison; the 95% CI will also be reported. Treatment group comparisons at Week 12 and all other visits will be tested.

The Fisher exact test will be used to perform treatment comparisons for AEs, discontinuations, and other categorical safety data among treatment groups. Change from baseline in continuous vital signs, physical characteristics, and other continuous safety variables, including laboratory variables, will be analyzed using an ANCOVA model with explanatory terms for treatment and the baseline value as a covariate. Type III sums of squares will be used. The significance of within-treatment group changes from baseline will be evaluated by testing whether the treatment group LS mean changes from baseline are different from zero using a t-statistic; the standard error for the LS mean change will also be displayed. Differences in LS means as compared to placebo will be displayed with the associated p-value and a 95% CI of the LS mean difference.

In addition to the LS means for each group, the within-group p-value, the standard deviation, minimum, median, and maximum will be displayed for continuous safety variables.

Data collected at early termination visits will be mapped to the next scheduled visit number for that patient.

The following is a description of the analysis groups planned for this study. Refer to Section 6.1.3 for the definitions of analysis populations used for the efficacy and safety analyses.

Analysis Groups for the Efficacy and Health Outcomes Data

Summaries of efficacy and health outcomes data for the Week 0 to Week 12 period, the Week 0 to Week 24 period with data up to rescue, and the Week 0 to Week 52 period with data up to rescue/switch will be presented according to the following treatment groups:

- Placebo: patients randomized to placebo excluding data obtained after the receipt of baricitinib 4 mg, either as rescue therapy or according to the protocol-designated switch to baricitinib 4 mg,
- Baricitinib 4 mg: patients randomized to baricitinib 4 mg excluding data obtained after the receipt of rescue therapy.

Summaries of efficacy and health outcomes data for patients who are rescued/switched will be presented according to the following treatment groups:

- Placebo (rescued): patients randomized to placebo who were rescued on or before Week 24,
- Placebo (switched, not rescued): patients switched to baricitinib from placebo at Week 24 and not rescued after the initiation of baricitinib,
- Placebo (switched, rescued): patients switched to baricitinib from placebo at Week 24 and rescued after initiation of baricitinib,
- Baricitinib (rescued): patients randomized to baricitinib who were rescued.

Analysis Groups for the Safety Data

Summaries of safety data for the Week 0 to Week 12 period and the Week 0 to Week 24 period with data up to rescue will be presented according to the following treatment groups:

- Baricitinib 4 mg: patients randomized to baricitinib excluding data obtained after the receipt of rescue therapy,
- Placebo: patients randomized to placebo using data obtained prior to the receipt of baricitinib as rescue therapy.

Summaries of safety data for the Week 0 to Week 52 period with data up to rescue/switch will be presented for baricitinib 4 mg for patients who were randomized to baricitinib excluding data obtained after receipt of rescue therapy.

Summaries of safety data for the Week 0 to Week 52 period (including data after rescue/switch), considering all baricitinib exposures, will be presented according to the following groups:

- Placebo (rescued): patients randomized to placebo who were rescued on or before Week 24 (using data collected after initiation of baricitinib),
- Placebo (switched, not rescued): patients switched to baricitinib from placebo at Week 24 and not rescued after initiation of baricitinib (using data collected after initiation of baricitinib),
- Placebo (switched, rescued): patients switched to baricitinib from placebo at Week 24 and rescued after initiation of baricitinib (using data collected after initiation of baricitinib),
- Baricitinib (rescued): patients randomized to baricitinib who were rescued (using data collected after initiation of baricitinib),
- All baricitinib exposures across all periods, regardless of initial treatment, use of rescue treatment, or protocol-defined switch of assigned treatment, ie, pooled across all cohorts who received baricitinib (using data collected after initiation of baricitinib).

Patients who have impaired renal function who are randomized to baricitinib 4 mg will be given a 2 mg dose but analyzed in the baricitinib 4 mg group. Additional analyses for selected efficacy and safety endpoints will use this subgroup of patients (see Section 6.14 [Safety Analyses] and Section 6.15 [Subgroup Analyses]).

Summaries of efficacy and safety data collected during the post-treatment follow-up period (Part C) will be presented separately from the summaries for Parts A and B and according to the following cohorts:

- Placebo (not rescued): patients randomized to placebo who discontinued on or before Week 24 and had never been rescued at the time of discontinuation,
- Placebo (rescued): patients randomized to placebo who were rescued on or before Week 24,
- Placebo (switched, not rescued): patients switched to baricitinib from placebo and not rescued after initiation of baricitinib,
- Placebo (switched, rescued): patients switched to baricitinib from placebo and rescued after initiation of baricitinib,
- Baricitinib (not rescued): patients randomized to baricitinib who were never rescued,
- Baricitinib (rescued): patients randomized to baricitinib who were rescued,
- All baricitinib exposures across all periods, regardless of initial treatment, use of rescue treatment, or protocol-defined switch of assigned treatment, ie, pooled across all cohorts who received baricitinib prior to entering Part C.

6.1.1. *Definition of Baseline and Postbaseline Measures*

The baseline period is defined as Visit 1 through pre-dose (expected at Visit 2). The treatment period for the blinded placebo and active study drug period (Part A) starts after the first study drug administration at Visit 2 (Week 0) and ends on the date of Visit 11 (Week 24) or an early discontinuation visit before Visit 11 (Week 24). The treatment period for the unblinded active study drug period (Part B) starts after the first study drug administration at Visit 11 (Week 24) and ends on the date of Visit 15 (Week 52) or an early discontinuation visit in the treatment period. The post-treatment follow-up period (Part C) starts after the date of Visit 15 (Week 52) or an early discontinuation visit and lasts until the follow-up visit, which occurs approximately 28 days following the last dose of the study drug, for patients who do not enter long-term extension Study JADY.

The baseline value for the efficacy analyses of the treatment period, including post-rescue analyses, is defined as the last non-missing measurement on or prior to the date of first study drug administration (expected at Week 0, Visit 2), regardless of rescue status. If the Physician's Global Assessment of Disease Activity score is recorded ≤5 days after the date of first study drug administration, it will still be used as the baseline value. Postbaseline measurements are collected after study drug administration (expected at Week 0, Visit 2) through Week 52 (Visit 15) or early discontinuation visit.

For data collected in the ePRO tablet related to efficacy assessments, unscheduled postbaseline visits that fall within the visit windows defined by Lilly will be summarized in the by-visit analyses if there is no scheduled visit available. That is, a +4-day window is applied to Visit 3 (Week 1) and Visit 4 (Week 2), and a +7-day window is applied to Visits 5 through 15 (Weeks 4 through 52) within the same treatment period (and while receiving the same treatment). For example, data should not be pulled across study periods for placebo group, and data cannot be pulled from postrescue visits to previous nonrescued visits. If there is more than 1 unscheduled visit within the defined visit window and no scheduled visit available, the unscheduled visit closest to the scheduled visit date will be used.

The baseline value for the safety analysis of the treatment period until rescue, protocol-defined switch, or end of study (if patients are not rescued to baricitinib 4 mg) is defined as the last non-missing measurement on or prior to the date of first study drug administration (expected at Week 0, Visit 2). The baseline value for the safety postrescue/switch period for patients randomized to placebo who are rescued/switched is defined as the last non-missing measurement obtained prior to initiation of the new treatment (baricitinib). The baseline value for the safety analyses of the postrescue period for patients randomized to baricitinib 4 mg who are rescued is the same as their treatment-period baseline. Postbaseline measurements for rescued/switched patients are collected after the rescue/switch visit through Week 52 (Visit 15) or an early discontinuation visit.

The baseline value for the efficacy analysis of the post-treatment follow-up period is defined as the last non-missing assessment at the time of Week 52 (Visit 15) or early discontinuation visit.

The baseline value for the safety analysis of the post-treatment follow-up period is defined as the last non-missing assessment on or prior to Week 52 (Visit 15) or early discontinuation visit.

6.1.2. *Derived Data*

- Age (year) = (number of months between informed consent date and July of birth year) / 12, and then truncated to a whole-year (integer) age

- Age group (<65 years old, ≥65 years old)
- Body mass index (BMI) (kg/m^2) = Weight (kg)/(Height (cm)/100)^2
- Duration of RA from onset (years) = [(Date of informed consent – Date of RA onset) + 1] / 365.25

  Date of RA onset is defined as the date when the patient had symptoms of RA. If year of onset is missing, duration of RA from onset will be set as missing. Otherwise, unknown month will be taken as January, and unknown day will be taken as 01. The duration of RA from onset will be rounded to 1 decimal place.

- Duration of RA from diagnosis (years) = [(Date of informed consent – Date of RA diagnosis) + 1] / 365.25

  Date of RA diagnosis is defined as the date when the patient was diagnosed with RA. See Duration of RA from onset (years) for the handling of missing year/month/date.

- Duration from onset of RA (years): 0 to <1; ≥1 to <5; ≥5
- Change from baseline = postbaseline measurement at Visit x – baseline measurement
- Percentage change from baseline at Visit x:
  - ((Postbaseline measurement at Visit x – baseline measurement) / baseline measurement)x100
- Baseline disease activity based on the DAS28 (calculated separately for DAS28-ESR and DAS28-hsCRP)
  - Low disease activity: DAS28 ≤3.2,
  - Moderate disease activity: DAS28 >3.2 to ≤5.1,
  - High disease activity: DAS28 >5.1

6.1.3. *Analysis Populations*

Modified Intent-to-Treat (mITT) Population: The mITT population will include all randomized patients who received at least 1 dose of study drug. The intention-to-treat principle remains preserved since the decision of whether or not to begin treatment could not be influenced by knowledge of the assigned treatment in this double-blind study. Patients will be analyzed according to the study drug to which they were randomized or assigned per protocol.

The efficacy and health outcomes analyses will be conducted on a mITT basis unless otherwise specified.

Analysis of structural progression (van der Heijde mTSS) will be conducted on a subset of the mITT population that includes all randomized patients who received at least 1 dose of the study drug and have non-missing baseline and at least 1 non-missing postbaseline measurement.

Per-Protocol (PP) Population: The PP population will include all mITT patients who are not deemed noncompliant with treatment in Part A, who do not have selected important protocol deviations (refer to Section 6.5 for more details) in Part A, and whose investigator site does not have significant good clinical practice (GCP) issues that require a report to regulatory agencies (regardless of study part). Qualifications and identification of the specific selected important protocol deviations that result in exclusion from the PP subsets will be determined prior to unblinding.

Since the primary and secondary efficacy endpoint analyses plus mTSS will be repeated using the PP population at Week 12 only (except that mTSS will be analyzed at Week 24), time constraints will be applied to the important protocol deviations that exclude a patient from the per protocol analysis population. For example, for treatment non-compliance, only patients with <80% compliance rate in the first 12 week period will be excluded from the PP population.

Safety Population: The safety population is defined as all randomized patients who received at least 1 dose of study drug and who did not discontinue from the study for the reason 'Lost to Follow-up' at the first postbaseline visit. Patients will be analyzed according to the analysis groups defined in Section 6.1.

Safety analyses up to the end of Part B will be conducted on this safety population.

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

Safety Follow-up Population: The safety follow-up population is defined as patients in the safety population who entered the follow-up period and who did not discontinue in the follow-up period for the reason 'Lost to Follow-up'.

6.1.4. *Handling of X-ray Data*

X-rays are scheduled at the following visits: Screening (Visit 1), Week 16 (Visit 9), Week 24 (Visit 11), and Week 52 (Visit 15). X-rays are obtained at an early termination only if the most recent x-ray was more than 12 weeks ago.

There may be instances in which the x-rays are obtained on a date different from the scheduled InForm visit date. Also, there may be instances in which the set of x-rays for a visit (2 hands/wrists, 2 feet) are split across dates, resulting from the need to re-acquire some images in the set. To maximize the use of the observed data and avoid imputation, the following rules will be applied in the analysis:

1. Postbaseline x-rays administered within 14 days prior to the scheduled InForm visit date, or within 30 days after the scheduled InForm visit date, will be associated with that respective InForm visit for analysis.
2. An x-ray set split across different dates will have the data combined as long as the images are collected within 30 days of each other. The earlier x-ray date will be considered to be the date of the x-ray set.

Any linear extrapolation will be based on the date of the x-ray set instead of the InForm visit date.

6.2. Adjustments for Covariates

The covariates used in the parametric statistical model (ANCOVA) of continuous data generally will include the parameter value at baseline. Inclusion of baseline in the ANCOVA model provides estimated adjusted (least squares) means for each treatment reflecting averages for the average baseline patient. Other independent variables that will be included in the models for efficacy and health outcomes analyses to adjust estimates of treatment effect include the stratifying variables country and joint erosion status. When the patient number in each country is deemed small (ie, n < 30) , countries will be combined into regions (ie, East Asia and Central America). The independent variables used in the logistic regression models will include country/region and joint erosion status (refer to Section 6.1). The stratification variables used in the analyses will be derived based on electronic case report form (eCRF) data and x-ray data collected by Bioclinica.

6.3. Handling of Dropouts or Missing Data

6.3.1. *Missing Data Imputation for Efficacy and Health Outcomes Endpoints*

In accordance with precedent set with other Phase 3 RA trials (Keystone et al. 2004, 2008, 2009; Cohen et al. 2006; Smolen et al. 2008, 2009), the following methods for imputation of missing data will be used when appropriate:

1. Non-responder imputation

All patients who discontinue the study or permanently discontinue the study treatment at any time for any reason will be defined as nonresponders for the NRI analysis for categorical variables, such as ACR20/ACR50/ACR70, from the time of discontinuation and onward. If, at the time of study database lock, there are patients with unresolved interruptions of study drug (ie, not yet re-started, to be confirmed as a temporary interruption, and not yet permanently discontinued), these patients will be treated as having been permanently discontinued from study drug for the purposes of data imputation for analysis purposes. Patients who are randomized to baricitinib 4 mg and who receive rescue therapy starting from Week 16 will be analyzed as nonresponders after the rescue visit and onward. Patients who are randomized to placebo and who receive rescue therapy starting from Week 16 will be analyzed as nonresponders after the rescue visit and on or prior to Week 24. Randomized patients without available data at a postbaseline visit will be defined as nonresponders for the NRI analysis at that visit.

2. Modified baseline observation carried forward

The mBOCF method will be used for the analysis of secondary continuous endpoints (unless otherwise stated). For patients who discontinue the study or permanently discontinue the study treatment because of an AE, including death and abnormal lab results reported as AEs, the baseline observation will be carried forward to subsequent time points for evaluation. For patients who discontinue the study or permanently discontinue the study treatment for reason(s) other than an AE, the last nonmissing postbaseline observation prior to discontinuation will be carried forward to the subsequent time points for evaluation. If, at the time of study database lock, there are patients with unresolved interruptions of study drug (ie, not yet re-started, to be confirmed as a temporary interruption, and not yet permanently discontinued), these patients will be treated as having been permanently discontinued from study drug for the purposes of data imputation for analysis purposes. For patients who are randomized to baricitinib 4 mg and who receive rescue therapy starting from Week 16, the last nonmissing observation at or before rescue will be carried forward to subsequent time points for evaluation. For patients who are randomized to placebo and who receive rescue therapy starting from Week 16, the last nonmissing observation at or before rescue will be carried forward to subsequent time points on or prior to Week 24 for evaluation. For patients whose baseline value is missing or who have no change from baseline values after the aforementioned imputation, then the change from baseline will take a value of "zero" indicating no improvement from baseline.

3. Modified last observation carried forward

For all continuous measures including safety analyses, the modified last observation carried forward (mLOCF) method will be a general approach to impute missing data unless otherwise specified. For patients who are randomized to baricitinib 4 mg and who receive rescue therapy starting from Week 16, the last nonmissing observation at or before rescue will be carried forward to subsequent time points for evaluation. For patients who are randomized to placebo and who receive rescue therapy starting from Week 16, the last nonmissing observation at or before rescue will be carried forward to subsequent time points on or prior to Week 24 for evaluation. For all other patients who discontinue from the study or permanently discontinue the study treatment for any reason, the last nonmissing postbaseline observation before discontinuation will be carried forward to subsequent time points for evaluation. If, at the time of study database lock, there are patients with unresolved interruptions of study drug (ie, not yet re-started, to be confirmed as a temporary interruption, and not yet permanently discontinued), these patients will be treated as having been permanently discontinued from study drug for the purposes of data imputation for analysis purposes. If a patient does not have a nonmissing observed record (or one imputed by other means) for a postbaseline visit, the last postbaseline record prior to the missed visit will be used for this postbaseline visit. After mLOCF imputation, data from patients with nonmissing baseline and postbaseline observations will be included in the analyses.

The mLOCF method may be used as a sensitivity approach to the mBOCF method with respect to the secondary endpoints. For safety analyses of the Week 0 to Week 52 period which includes data collected after rescue, the last nonmissing assessment prior to a missed visit will be carried forward if the assessment occurred after the first dose of baricitinib (ie, after first study dose for groups originally randomized to baricitinib 4 mg, after first rescue dose for patients randomized to placebo who are later rescued, and after first baricitinib dose for patients randomized to placebo who switch treatment at Week 24 according to the protocol).

6.3.2. *Missing Data Imputation for the Structural Progression Endpoint (van der Heijde mTSS)*

1. Primary method: Linear extrapolation method

The primary approach to the analysis of the structural progression data will be the linear extrapolation method, which assumes that individual patients would continue to accrue structural damage in their joints at the same rate that was observed at their time of last observation. The linear extrapolation method has been established as an appropriate missing data imputation method in other Phase 3 RA trials (Keystone et al. 2004, 2008, 2009; Cohen et al. 2006; Smolen et al. 2008, 2009). The linear extrapolation method will be used for analysis of the structural progression endpoint (van der Heijde mTSS), including bone erosion scores and joint space narrowing, to impute missing data at time points when analyses are conducted (including the analyses at Week 24 and Week 52). For patients who discontinue from the study or for patients who miss a radiograph for any reason, baseline data and the most recent radiographic data prior to discontinuation or prior to the missed radiograph, adjusted for time, will be used for linear extrapolation to impute missing data at subsequent scheduled time points, including the primary analysis at Week 24. For patients who receive rescue therapy starting at Week 16 or at any time point thereafter, baseline data and the most recent radiographic data prior to initiation of rescue therapy, adjusted for time, will be used for linear extrapolation. For patients who discontinue from the study or permanently discontinue the study treatment after receiving rescue therapy, the radiographic data collected prior to initiation of rescue therapy should supersede the data collected before discontinuation.

All patients originally randomized to placebo will be switched to active treatment at Week 24; thus, there will be no observed data for the placebo arm after Week 24. Therefore, baseline data and the most recent radiographic data prior to initiation of the new therapy, adjusted for time, will be used for linear extrapolation at Week 52. Missing postbaseline values will be imputed only if both a baseline value and a postbaseline value from another time point are available, as long as the patient was receiving the same treatment at each applicable time point. Missing postbaseline values will not be imputed using postbaseline data from later time points.

For example:

(1) For a baricitinib-treated patient who did not require rescue therapy:

- Missing data at Week 16 will be extrapolated from baseline and an early discontinuation visit prior to Week 16, if available.
- Missing data at Week 24 will be extrapolated from baseline and the last available postbaseline value prior to Week 24.
- Missing data at Week 52 will be extrapolated from baseline and the last available postbaseline value prior to Week 52, if available.

(2) For a baricitinib- or placebo-treated patient who was rescued at Week 16:

- Missing data at Weeks 24 or 52 will be extrapolated from baseline and Week 16.

(3) For a placebo-treated patient who did not require rescue therapy:

- Missing data at Week 16 will be extrapolated from baseline and an early discontinuation visit prior to Week 16, if available.

- Missing data at Week 24 will be extrapolated from baseline and the last available postbaseline value prior to Week 24.
- Missing data at Week 52 will be extrapolated from baseline and the last available postbaseline value at or prior to Week 24.

2. Other method: Last observation carried forward

A last-observation-carried-forward (LOCF) approach will also be used for the Week 24 analyses of mTSS, bone erosion scores, and joint space narrowing. Data observed at Week 24 will be analyzed according to randomized treatment group regardless of rescue status. Patients with missing data at Week 24, as a result of study discontinuation or unreadable x-rays, will have their data imputed by last postbaseline observation carried forward.

Details of the missing data imputation rules, methods for between-reader adjudication, score derivations, and other details for the bone erosion scores, joint space narrowing scores, and mTSS are provided in Appendix 2 (van der Heijde Modified Total Sharp Score) and Appendix 4.

Table 6-1 summarizes the imputation techniques for the various endpoints.

Table 6-1 Imputation Techniques for Various Efficacy and Safety Variables

Endpoint	Imputation Method
ACR 20/50/70 responses, DAS28-hsCRP and DAS28-ESR LDA and remission, EULAR response criteria based on the 28-joint DAS, ACR/EULAR remission according to the Boolean-based definition, HAQ-DI improvement from baseline ≥0.22 and ≥0.3, CDAI LDA and remission, and SDAI LDA and remission	NRI
JSN, bone erosion scores, and mTSS	Linear extrapolation method and LOCF
mTSS ≤0, mTSS ≤0.5, and mTSS ≤SDC	NRI
HAQ-DI, DAS28-hsCRP	mBOCF/mLOCF
Duration and severity of morning joint stiffness, worst tiredness, and worst joint pain assessed in the diaries	See Section 6.13
Continuous efficacy, health outcome, and safety measures (laboratory values, vital signs, etc.)	mLOCF

Abbreviations: ACR = American College of Rheumatology; CDAI = Clinical Disease Activity Index; DAS28 = Disease Activity Score modified to include the 28 diarthrodial joint count; EULAR = European League Against Rheumatism; HAQ-DI = Health Assessment Questionnaire–Disability Index; hsCRP = high-sensitivity C-reactive protein; JSN = joint space narrowing; LDA = low disease activity; mBOCF = modified baseline observation carried forward; mLOCF = modified last observation carried forward; LOCF = last observation carried forward; mTSS = modified Total Sharp Score; NRI = nonresponder imputation; SDAI = Simplified Disease Activity Index; SDC = smallest detectable change.

6.4. Patient Disposition

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

Patient disposition will be summarized using the mITT population. Frequency counts and percentages of patients who complete through Week 24 or discontinue early from the study on or prior to Week 24 will be summarized separately by treatment group for patients who are not rescued and for patients who are rescued, along with their reason for study discontinuation. Frequency counts and percentages of patients who complete through Week 52 or discontinue early from the study after Week 24 and on or prior to Week 52 will be summarized separately by treatment group for patients who are not rescued/switched and for patients who are rescued/switched, along with their reason for study discontinuation. Among patients who completed the study, frequency counts and percentages of patients who entered Study JADY, completed the post-treatment follow-up visit, or did not complete the follow-up visit will be presented by treatment group. Among patients who discontinued early from the study, frequency counts and percentages of patients who completed the post-treatment follow-up visit or did not complete the follow-up visit will be presented by treatment group. Reasons for not completing the post-treatment follow-up visit will also be presented.

A listing of patients who screen-failed and a listing of disposition of randomized patients will be provided with the extent of their participation in the study and the reason for screen failure or discontinuation, respectively. A listing of all randomized patients will also be provided.

6.5. Protocol Deviations

Protocol deviations will be tracked by the clinical team, and their importance will be assessed by key team members during protocol deviation review meetings. Out of all important protocol deviations identified, a subset occurring during Part A prior to the primary endpoint (Week 12) with the potential to affect efficacy analyses will result in exclusion from the PP population.

Potential examples of deviations include patients who receive excluded concomitant antirheumatic 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.

Important protocol deviations are the following (those indicated with an asterisk [*] may exclude a patient from the PP analysis population):

- Protocol inclusion or exclusion criteria:
    - General
        - Are unable to read, understand and give written informed consent
        - Are pregnant or nursing at the time of study entry (*)
    - Prior infection or comorbidity
        - Have a history of lymphoproliferative disease; or have signs or symptoms suggestive of possible lymphoproliferative disease, including lymphadenopathy or splenomegaly; have active primary or recurrent

malignant disease; or have been in remission from clinically significant malignancy for <5 years

  - Have been exposed to a live vaccine within 12 weeks prior to planned randomization
  - Have been exposed to herpes zoster vaccination within 30 days of planned randomization
  - Have a current or recent (30 days prior to study entry) clinically serious viral, bacterial, fungal, or parasitic infection
  - Have had symptomatic herpes zoster infection within 12 weeks prior to study entry
  - Have a history of disseminated/complicated herpes zoster (multi dermatomal involvement, ophthalmic zoster, central nervous system involvement, or post therapeutic neuralgia)
  - Have a history of active hepatitis B virus (HBV), hepatitis C virus (HCV), or human immunodeficiency virus (HIV)
  - Have had household contact with a person with active tuberculosis (TB) and did not receive appropriate and documented prophylaxis for TB
  - Have evidence of active TB or have previously had evidence of active TB and did not receive appropriate and documented treatment
  - Have a diagnosis of Felty's syndrome

- Current infection
  - Have evidence of active TB as documented by a purified protein derivative (PPD) test (≥5-mm induration between approximately 2 and 3 days after application, regardless of vaccination history), medical history, clinical symptoms, and abnormal chest x-ray at screening (*)
  - Have evidence of latent TB (as documented by a positive PPD, no clinical symptoms consistent with active TB and a normal chest x-ray at screening) unless patients complete at least 4 weeks of appropriate treatment prior to randomization and agree to complete the remainder of treatment while in the trial
  - Have a positive test for HBV
  - Have HCV (positive for anti-hepatitis C antibody with confirmed presence of HCV)
  - Have evidence of HIV infection or positive HIV antibodies

- Medications

    - Have not had regular use of MTX for at least 12 weeks prior to study entry. The dose of MTX must have been a stable, unchanging oral dose of 7.5 to 25 mg/week (or the equivalent injectable dose) for at least 8 weeks prior to study entry.
    - Have received conventional disease-modifying antirheumatic drugs (cDMARDs) (eg, gold salts, cyclosporine, azathioprine, or any other immunosuppressives) other than MTX, hydroxychloroquine (up to 400 mg/day), or sulfasalazine (up to 3000 mg/day) within 4 weeks prior to study entry at a stable dose for at least 8 weeks prior to study entry; if either was recently discontinued, the patient must not have taken any dose within 4 weeks prior to study entry. Immunosuppression related to organ transplantation was not permitted.
    - Have ever received any bDMARD (such as TNF, interleukin-1 [IL-1], IL-6, or T-cell- or B-cell-targeted therapies)
    - Have been receiving an unstable dosing regimen of oral, intramuscular, or intravenous corticosteroids within 6 weeks of randomization
    - Are currently receiving corticosteroids at doses >10-mg per day of prednisone (or equivalent) at study entry
    - Have started NSAID treatment for signs and symptoms of RA within 6 weeks of randomization
    - Have been receiving an unstable dosing regimen of NSAIDs within 6 weeks of randomization
    - Have received interferon therapy (such as Roferon-A, Intron-A, Rebetron, Alferon-N, Peg-Intron, Avonex, Betaseron, Infergen, Actimmune, or Pegasys) within 4 weeks prior to study entry or are anticipated to require interferon therapy during the study
- Have any specific abnormalities on screening laboratory tests
    - Have aspartate aminotransferase (AST) >1.5 times (1.5 ×) the upper limit of normal (ULN) on screening laboratory testing (1 retest allowed within approximately 2 weeks)
    - Have alanine aminotransferase (ALT) >1.5 × ULN on screening laboratory testing (1 retest allowed within approximately 2 weeks)
    - Have total bilirubin ≥1.5 × ULN on screening laboratory testing (1 retest allowed within approximately 2 weeks)
    - Have hemoglobin <10.0 g/dL on screening laboratory testing (1 retest allowed within approximately 2 weeks)

        - Have total white blood cell count <2500 cells/µL on screening laboratory testing (1 retest allowed within approximately 2 weeks)
        - Have absolute neutrophil count <1200 cells/µL on screening laboratory testing (1 retest allowed within approximately 2 weeks)
        - Have lymphocyte count <750 cells/µL on screening laboratory testing (1 retest allowed within approximately 2 weeks)
        - Have platelets <100,000 cells/µL on screening laboratory testing (1 retest allowed within approximately 2 weeks)
        - Have eGFR <40 mL/min/1.73 m^2 on screening laboratory testing (1 retest allowed within approximately 2 weeks)

- Use of prohibited concomitant medication during study
    - Use of any new or any change of dose in steroids in the absence of an AE or prior to rescue. Use for treatment of RA disease activity will be considered an important protocol deviation (*)
    - Use of any new or increased analgesics or NSAIDs in the absence of an AE or prior to rescue
    - Change in cDMARD therapy in the absence of an AE. Change for treatment of RA disease activity will be considered an important protocol deviation (*)
    - Use of any biologic or interferon therapy (*)
- Study drug administration and compliance
    - Drug dispensing error (e.g. patient dispensed wrong package and received incorrect dose or drug) (*)
    - Patient received drug that was declared “Not Fit for Use” (*)
    - Use of expired clinical trial Material (*)
    - Significant non-compliance as defined per protocol (*)
    - IWRS data entry errors that impact patient stratification at randomization
- Protocol non-compliance
    - Patient met temporary or permanent study drug discontinuation criteria and study drug not discontinued per protocol
    - Inadvertent unblinding (*)
    - Fraud (*)
    - Other significant GCP issues (*)

A summary of the number and percentage of patients with an important protocol deviation by treatment group, overall, and by type of deviation will be provided. Although the PP population is for the primary endpoint (Week 12) in Part A only, important protocol deviations will be summarized for Parts A and B. Individual patient listings of important protocol deviations will be provided. A summary of reasons patients were excluded from the PP population will be provided by treatment group, and a patient listing will also be generated.

6.6. Patient Characteristics

Patient characteristics including demographics will be summarized using the mITT analysis population by treatment group. The summary will include descriptive statistics such as the number of patients (n), mean, standard deviation (SD), median, minimum, and maximum for continuous measures and frequency counts and percentages for categorical measures. No formal statistical comparisons will be made among treatment groups, unless otherwise stated.

6.6.1. *Demographics*

Patient demographics will be summarized by treatment group using the mITT population. The following demographic information will be presented for this study.

- Age and Age group
- Gender
- Race
- Region/Country
- Weight
- Height
- BMI
- Waist circumference
- Habits (alcohol, smoking, caffeine)

In addition, patient demographics will be summarized by region/country using the mITT population. A listing of patient demographics will be produced.

6.6.2. *Baseline Disease Characteristics*

Baseline disease characteristics will be summarized by treatment group using the mITT population. The following baseline disease information (but not limited to only these) will be categorized and presented for baseline cardiovascular and renal comorbidities, baseline RA clinical characteristics, baseline health outcomes measures, and other baseline demographic and disease characteristics:

- Duration from onset of RA (years),
- Duration from onset of RA category (<1 years, ≥1 to <5, ≥5 years)
- Years since RA Diagnosis
- Patient's assessment of pain (by visual analog scale [VAS] in mm)
- Patient's Global Assessment of Disease Activity (by VAS in mm)
- Physician's Global Assessment of Disease Activity (by VAS in mm)
- HAQ-DI total score
- ESR (mm/hr)
- hsCRP (mg/L)

- DAS28-ESR and DAS28-hsCRP
- DAS28-ESR and DAS28-hsCRP category (≤3.2, >3.2 to ≤5.1, >5.1)
- Anti-citrullinated peptide antibody (ACPA) positivity and rheumatoid factor (RF) positivity status
- Number of tender joints out of 28 (TJC28)
- Number of swollen joints out of 28 (SJC28)
- Number of tender joints out of 68 (TJC68)
- Number of swollen joints out of 66 (SJC66)
- eGFR
  - eGFR <60 mL/min/1.73 m^2
  - eGFR ≥60 mL/min/1.73 m^2
- MTX dose (mg/week)
- Number of cDMARDs currently used (0, 1, ≥2)
- Number of cDMARDs previously used (0, 1, 2, ≥3)
- Type of cDMARDs currently used
  - Use of MTX only
  - Use of MTX + other cDMARDs
- Use of oral corticosteroid and dose (mg/day) expressed as prednisone-equivalent dose
- Use of statin therapy (Yes/No)

In addition, the baseline RA clinical characteristics will be summarized by country/region using the mITT population. A listing of patient demographics and selected baseline characteristics will be produced.

6.7. Previous and Concomitant Medications

Summaries of previous and concomitant medications used for RA including cDMARDs (MTX only, MTX and other cDMARDs; see Section 6.6) will be based on the safety population.

At screening, medications used for RA that had been previously discontinued are recorded for each patient. A summary of previous medications used for RA will be prepared using frequency counts and percentages by World Health Organization (WHO) drug anatomical therapeutic chemical (ATC) classification and treatment group.

Concomitant therapy will be recorded at each visit and will be classified according to the WHO drug dictionary and summarized using frequency counts and percentages by WHO drug ATC classification sorted by decreasing frequency and treatment group.

Concomitant medications are defined as:

- Concomitant medications with a stop date on or after the date of the first dose of the study drug and that precedes Week 24 (Visit 11) or an early discontinuation visit prior to Week 24 will be included in the concomitant medications summaries for the analysis at the end of Part A. 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 Part A.

- For patients who continue in Part B, concomitant medications with a stop date that is after Week 24 or ongoing will be included in the concomitant medications summaries at the end of Part B.

Summaries of concomitant medications will be provided using data up to rescue/switch for the following categories:

- concomitant medications used for RA
- concomitant medications excluding RA and statin therapies
- concomitant medications - statins

Concomitant therapies will also be summarized for the all baricitinib exposures group including data after rescue/switch.

6.8. Historical Diagnoses and Pre-Existing Conditions

The number and percentage of patients with selected historical diagnoses, as listed on the eCRF, will be summarized by treatment group using the mITT population. Historical diagnoses will be identified using the Medical Dictionary for Regulatory Activities (MedDRA®, most current available version) algorithmic standardized MedDRA queries (SMQs) or similar Lilly-defined lists of preferred terms (PTs) of interest.

Pre-existing conditions are defined as those conditions recorded on the Pre-existing Conditions and AEs eCRF page with a start date prior to the date of informed consent. Note that conditions with a partial or missing start date 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 identified using the MedDRA SMQs or similar Lilly-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 mITT population.

6.9. Treatment Compliance

Patient compliance with study medication will be assessed from randomization visit (Week 0) to Visit 15 (Week 52) during the treatment period. Compliance will be summarized by treatment groups from randomization to visit of rescue/switch using the mITT population.

All patients are expected to be taking1 tablet of baricitinib or baricitinib placebo daily in Parts A and B regardless of rescue; a patient is considered significantly noncompliant if he or she misses >20% of the prescribed doses during the study. 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 study drug. For patients who have their dose interrupted because of safety reasons, the period of dose withholding will be taken into account in the compliance calculation as outlined below.

Compliance to baricitinib and baricitinib placebo in the period of interest will be calculated as follows:

$$\frac{\text{actual total \# of tablets used}}{\text{Expected total \# of tablets used}} \text{ x } 100$$

Where:

the expected total # tablets used = the number of days in the period of interest;

the actual total # of tablets used = total number dispensed – total number returned in the period.

If a patient has dose interrupted for safety reasons during the period, the total number of days drug was withheld will be deducted from the expected total # of tablets used.

Patients who are significantly noncompliant in Part A will be excluded from the PP population.

The summary statistics of percent compliance and non-compliance rate will be summarized by treatment group. The sub-total percent compliance for Weeks 0 through 12, the sub-total percent compliance for Weeks 0 through 24, and grand-total percent compliance for Weeks 0 through 52 with data up to the time of rescue/switch will be presented, as well, along with the associated non-compliance rates. The number of expected doses, tablets dispensed, tablets returned, and percent compliance will be listed by patient.

6.10. Treatment Exposure

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

Overall exposure will be summarized in total patient-years which are calculated according to the following:

> Exposure in patient-years = sum of duration of exposure in days (for all patients in treatment group) / 365.25.

The summary for the first database lock (end of Part A) will include exposure to each treatment plus exposure for the subset of patients who are rescued during Part A from each randomized therapy.

The summary for the final lock (end of Parts A and B) will include all exposures for patients randomized to baricitinib and placebo, plus all exposures subsequent to rescue or switch to baricitinib. For the baricitinib treatment group, exposure will be calculated including the subset of patients who are rescued.

Duration of exposure will be calculated as follows:

1. Baricitinib group:
   - Duration of exposure to baricitinib excluding exposure after the initiation of rescue therapy will be calculated as: [date of last study drug treatment prior to rescue – date of first study drug dose (Visit 2) + 1].
   - Duration of exposure to baricitinib after the initiation of rescue therapy will be calculated as: (date of last study drug treatment – date of first baricitinib dose as rescue therapy + 1).

2. Placebo group:
    - Duration of exposure to placebo excluding exposure after the initiation of rescue therapy will be calculated as: [date of last study drug treatment prior to rescue – date of first study drug dose (Visit 2) + 1].
    - Duration of exposure to baricitinib after the initiation of rescue therapy will be calculated as: [date of last study drug treatment – date of first baricitinib dose as rescue therapy + 1].
    - Duration of exposure to baricitinib after switch will be calculated as: (date of last study drug treatment (before rescue, if applicable) – date of first dose of study drug after switch + 1).

Summaries of treatment exposure will also include all baricitinib exposures.

6.11. Efficacy Analyses

Efficacy analyses will use the mITT population defined in Section 6.1.3 and will generally be analyzed according to the following 2 formats:

- Week 0 (Visit 2) to Week 52 (Visit 15) with data included up to the time of rescue/switch. Formal statistical tests will be used to compare the baricitinib group to the placebo group for these efficacy outcomes through Week 24;
- Week 0 (Visit 2) to Week 52 (Visit 15) for rescued/switched patients with data included after rescue/switch. These efficacy results will be summarized by treatment group, but no formal statistical comparisons will be performed.

6.11.1. Primary Outcome and Methodology

The primary efficacy measure will be the proportion of patients achieving ACR20 at Week 12 based on the change in ACR response values from baseline. ACR20 is defined as at least 20% improvement from baseline in the following ACR Core Set variables:

- TJC68
- SJC66
- An improvement of ≥20% from baseline in at least 3 of the following 5 assessments:
    - Patient's assessment of pain (VAS)
    - Patient's Global Assessment of Disease Activity (VAS)
    - Physician's Global Assessment of Disease Activity (VAS)
    - Patient's assessment of physical function as measured by the HAQ-DI
    - Acute phase reactant as measured by hsCRP

Refer to the algorithm to calculate ACR response in Appendix 1.

The primary efficacy analysis compares ACR20 at Week 12 between baricitinib and placebo using the mITT population. The PP population results will be used as supportive to the mITT results. A logistic regression model with treatment, country/region, and joint erosion status (1-2 joint erosions plus seropositivity versus at least 3 joint erosions) in the model will be used to test the treatment difference between baricitinib and placebo in the proportion of patients achieving ACR20 response at Week 12 using the Wald test at 2-sided significance level of 0.05. The NRI approach will be used to impute response status for patients with completely missing data. The p-value and 95% CI of odds ratio from the logistic regression model will be reported. The 95%

CI of responder rate difference will be calculated using the Newcombe-Wilson method (Newcombe 1998) without continuity correction. Treatment-by-country/region interaction will be added to the logistic regression model as a sensitivity analysis and results from this model will be compared to the primary model. If the treatment-by-country/region interaction is significant at a 2-sided 0.1-alpha level, the nature of this interaction will be inspected as to whether it is quantitative (ie, the treatment effect is consistent in direction but not size of effect) or qualitative (the treatment is beneficial in some but not all countries). If the treatment-by-country/region interaction effect is found to be quantitative, results from the primary model will be presented. If the treatment-by-country/region interaction effect is found to be qualitative, further inspection will be used to identify in which countries baricitinib is found to be more beneficial.

6.11.2. Secondary Efficacy Analyses

Unless otherwise stated, the following analyses will be performed to test for superiority of baricitinib relative to the placebo group.

The secondary efficacy measures are:

- Change from baseline to Week 12 in HAQ-DI score
- Change from baseline to Week 12 in DAS28-hsCRP
- Proportion of patients achieving an SDAI score ≤3.3 at Week 12
- Mean duration of morning joint stiffness (paper diary; 7 days prior to Week 12)
- Mean severity of morning joint stiffness (paper diary; 7 days prior to Week 12)
- Mean “Worst Tiredness” NRS (paper diary; 7 days prior to Week 12)
- Mean “Worst Joint Pain” NRS (paper diary; 7 days prior to Week 12)

The secondary efficacy measures will be analyzed using the mITT population. The PP population results will be used as supportive to the mITT results. These analyses are described as follows:

- **Comparison between baricitinib and placebo in change from baseline to Week 12 in HAQ-DI score and in DAS28-hsCRP**: an ANCOVA model with baseline score, country/region, joint erosion status, and treatment group in the model will be used to test the treatment difference between baricitinib and placebo in change from baseline to Week 12 in HAQ-DI and in DAS28-hsCRP. Type III sums of squares for the LS means will be used for the statistical comparison; 95% CI will also be reported. The LS mean difference, SE, CI, and p-value will be presented. The mBOCF approach described in Section 6.3.1 will be used to impute missing data.

- **Comparison between baricitinib and placebo in proportion of patients achieving an SDAI score ≤3.3 at Week 12**: A logistic regression model with treatment, country/region, and joint erosion status in the model will be used to test the treatment difference between baricitinib and placebo in the proportion of patients achieving an SDAI score ≤3.3 at Week 12. The NRI method as described in Section 6.3.1 will be used to impute missing data. The p-value and 95% CI of odds ratio from the logistic regression model will be reported for statistical inference. The 95% CI of responder rate difference will be calculated using the Newcombe-Wilson method (Newcombe 1998) without continuity correction.

- **Comparison between baricitinib and placebo in mean duration of morning joint stiffness in the 7 days prior to Week 12**: refer to Section 6.13 for the statistical methods used for the analysis of mean duration of morning joint stiffness.
- **Comparison between baricitinib and placebo in mean severity of morning joint stiffness in the 7 days prior to Week 12**: refer to Section 6.13 for the statistical methods used for the analysis of severity of morning joint stiffness.
- **Comparison between baricitinib and placebo in mean "Worst Tiredness" NRS in the 7 days prior to Week 12**: refer to Section 6.13 for the statistical methods used for the analysis of "worst tiredness."
- **Comparison between baricitinib and placebo in mean "Worst Joint Pain" NRS in the 7 days prior to Week 12**: refer to Section 6.13 for the statistical methods used for the analysis of "worst joint pain."

6.11.3. *Efficacy Analyses in Follow-Up Period*

Summaries of efficacy data collected during the post-treatment follow-up period (Part C) will be presented separately from the main summaries (Parts A and B). The efficacy measures to be summarized include DAS28-hsCRP, DAS28-ESR, ACR component scores (TJC68, SJC66, patient's assessment of pain, Patient's Global Assessment of Disease Activity, Physician's Global Assessment of Disease Activity, HAQ-DI, and hsCRP), CDAI, and SDAI. The summary statistics of these measures at study baseline, follow-up baseline, at follow-up visit, and for the changes from study baseline and from follow-up baseline to follow-up visit will be presented based on the groups defined in Section 6.1 for the follow-up population. The follow-up baseline is defined as the data collected at the last visit in Parts A and B prior to entering the follow-up period. Patients who have data for both the follow-up baseline and the follow-up visit are included in the analysis. Some other analyses for efficacy measures will be explored as well.

6.12. Pharmacokinetic/Pharmacodynamic Analysis (PK/PD)

No PK/PD samples are collected and no analyses are planned for this study.

6.13. Health Outcomes Analyses

6.13.1. Diary Data

6.13.1.1. Data Handling and Missing Data Imputation

The paper diary data (duration of morning joint stiffness, severity of morning joint stiffness, worst joint pain, worst tiredness, and recurrence of morning joint stiffness during the day) will be assessed daily from Week 0 (Visit 2) to Week 12 (Visit 7). The primary analysis is based on the scores collected in the 7 days prior to the Week 12 visit date. Similar analysis will be done for 7 days prior to each scheduled visit (Week 1, Week 2, Week 4, and Week 8) using the primary analysis method. The supportive analysis is based on daily data modeled by weekly means. The weeks are defined as follows: Day 1 on diary, Days 2-8 (Week 1), Days 9-15 (Week 2), and so on up to Days 79-85 (Week 12).

For the primary analysis, the average of the scores collected in the 7 days prior to Week 12 will be used. If there are fewer than 4 nonmissing scores, the 7-day window will be shifted back in time (toward baseline) 1 day at a time until there are 4 nonmissing scores available in the 7-day window. Then the average of the 4 scores will be used in the Week 12 analysis. This will be repeated for each scheduled visit (week 1, week 2, week 4, week 8) prior to week 12.

For patients who discontinue the study or permanently discontinue the study treatment, the scores collected in the last 7 days prior to discontinuation will be used in the analysis. If a patient undergoes a temporary interruption of study drug at the time of study database lock, this patient will be treated as having permanently discontinued from study drug in the missing data imputation. If there are fewer than 4 nonmissing scores, a similar approach will be used to find 4 nonmissing scores in a 7-day window and the average score will be used in the Week 12 analysis. Patients who do not have at least 4 scores in any postbaseline 7-day window up to Week 12 will have their Day 1 score carried forward to the endpoint for analysis. If the Day 1 value cannot be obtained, the first available postbaseline score will be used to impute the Day 1 score so that the patient can be included in the analysis.

For the analysis at each scheduled visit using the primary analysis method, the same imputation rule for the primary analysis will be used. The imputed value for missing Day 1 measures from postbaseline assessments as described in the primary analysis will not be used in the change from Day 1 analysis.

For the supportive analysis, if ≥4 scores during a defined week are missing, the daily scores collected in the week will not be used in the analysis as a protective measure against potentially unreliable or non-representative data.

6.13.1.2. Analysis of Diary Data

The effect of baricitinib compared to placebo up to Week 12 with regard to patient-reported outcomes (PROs) will be evaluated by the following measures:

- **Duration of morning joint stiffness**

  Duration of morning joint stiffness is a patient-administered item that allows for the patients to enter the length of time in minutes that their morning joint stiffness lasted each day (using a paper diary) through Visit 7 (Week 12). Duration of morning stiffness will be recorded in minutes. Durations recorded as longer than 12 hours (720 minutes) will be truncated to 720 minutes for analysis.

  A nonparametric method involving the Hodges-Lehmann estimator will be used to test the treatment difference in the average durations (of morning joint stiffness) collected in the 7 days prior to Week 12. For the comparison between the baricitinib group and the placebo group, the median score and its 95% CI from the Hodges-Lehmann estimator will be presented. The p-value for the median difference will be calculated using the Wilcoxon rank-sum test. A 95% CI for the median score within each treatment group will be calculated using distribution-free confidence limits. Missing data will be handled as described for the main analysis in Section 6.13.1.1. This analysis of duration of morning joint stiffness will be conducted using the mITT population. The PP population results will be used as supportive evidence to the mITT results. The primary analysis of duration of morning joint stiffness will also be repeated for each scheduled visit using the mITT population.

In addition, the average change of duration (of morning joint stiffness) from Day 1 across the 7 days prior to Week 12 on diary will be compared between the baricitinib group and the placebo group using the same analysis method described. The imputed value for missing Day 1 durations from postbaseline assessments as described in Section 6.13.1.1 will not be used in the change from Day 1 analysis.

- **Severity of morning joint stiffness NRS**

  Severity of morning joint stiffness NRS is a patient-administered, single-item, 11-point horizontal scale anchored at 0 and 10, with 0 representing "no joint stiffness" and 10 representing "joint stiffness as bad as you can imagine." Patients rate their morning joint stiffness each day (using a paper diary) up to Visit 7 (Week 12) by selecting 1 number that describes their overall level of joint stiffness at the time they wake up.

  An ANCOVA model will be used to test for treatment differences in the severity of morning joint stiffness NRS collected in the 7 days prior to Week 12, with treatment, country/region, and baseline joint erosion status (1-2 joint erosions plus seropositivity versus at least 3 joint erosions) as fixed effects and Day 1 value as a covariate. Missing data will be handled as described for the primary analysis in Section 6.13.1.1. Type III sums of squares for the LS means will be used for the statistical comparison; the 95% CI will also be reported. The LS mean difference, SE, CI, and p-value will be presented for the treatment difference between the baricitinib group and the placebo group. This analysis of severity of morning joint stiffness NRS will be conducted using the mITT population. The PP population results will be used as supportive evidence to the mITT results. The primary analysis of severity of morning joint stiffness NRS will also be repeated for each scheduled visit using the mITT population.

- **Tiredness severity NRS ("Worst Tiredness")**

  Worst tiredness NRS is a patient-administered, single-item, 11-point horizontal scale anchored at 0 and 10, with 0 representing "no tiredness" and 10 representing "as bad as you can imagine." Patients rate their tiredness each day (using a paper diary) through Visit 7 (Week 12) by selecting the 1 number that describes their worst level of tiredness during the past 24 hours.

  The primary analysis of "Worst tiredness" will be analyzed in a similar manner as described for severity of morning joint stiffness NRS. This primary analysis of "Worst tiredness" will be conducted using the mITT population. The PP population results will be used as supportive evidence to the mITT results. The primary analysis of "Worst tiredness" will also be repeated for each scheduled visit using the mITT population.

- **Pain severity NRS ("Worst Joint Pain")**

  Worst Joint Pain NRS is a patient-administered, single-item, 11-point horizontal scale anchored at 0 and 10, with 0 representing "no pain" and 10 representing "pain as bad as you can imagine." Patients rate their joint pain each day (using a paper diary) through Visit 7

(Week 12) by selecting the 1 number that describes their worst level of joint pain during the past 24 hours.

The primary analysis of "Worst joint pain" will be analyzed in a similar manner as described for severity of morning joint stiffness NRS. This primary analysis of "Worst joint pain" will be conducted using the mITT population. The PP population results will be used as supportive evidence to the mITT results

- **Recurrence of joint stiffness during the day**

  Recurrence of joint stiffness throughout the day is a patient-administered single-item question assessed each day (using a paper diary) through Visit 7 (Week 12) that asks "Did you have any joint stiffness throughout today, other than when you woke up?" (Yes/No).

  Recurrence of stiffness during the day will be summarized as a percentage of days with the event at weekly intervals (as defined in Section 6.13.1.1) up to Visit 7 (Week 12). Percentage of days will be analyzed by week using analysis of variance (ANOVA) with country/region, treatment group, and joint erosion status value in the model. Missing data imputation will follow what was described for the supportive analysis in Section 6.13.1.1.

6.14. Safety Analyses

Safety analyses for this study will be provided in the following manner:

- The analysis of safety parameters during Part A wherein the comparison between baricitinib and placebo is of primary interest, informing risks to treatment that can be evaluated relative to the potential efficacy advantages shown over the same acute treatment time course. The baricitinib-placebo differences will be formally statistically compared, where applicable. Data collected up to the time of rescue (if applicable) will be used in these comparisons.
- Secondly, the summary analysis of safety parameters which includes data collected during Part B will be provided. No formal statistical comparisons are planned.
- Thirdly, summarization of safety parameters after inception of treatment with baricitinib, experienced from the moment of randomization, from the time of switching from treatment with placebo to baricitinib, or from the point of rescue to treatment with baricitinib will be provided. As these comparisons do not result from some form of randomization, no formal statistical comparisons will be performed. These summaries will include all qualifying data from Parts A and B.

Thus, safety results will generally be prepared according to 4 formats:

- Part A, Week 0 (Visit 2) to Week 12 (Visit 9) covering exposures to randomized treatments prior to the primary timepoint at Week 12,
- Part A, Week 0 (Visit 2) to Week 24 (Visit 11) covering all of Part A but with censoring for patients after the point of rescue,
- Parts A and B combined from Week 0 (Visit 2) to Week 52 (Visit 15) for patients randomized to baricitinib but with censoring after the point of rescue,

- Parts A and B combined from Week 0 (Visit 2) to Week 52 (Visit 15) for patients randomized to baricitinib and for patients who change treatment to baricitinib from placebo covering all exposure to baricitinib.

Safety topics that will be addressed include the following: AEs, clinical laboratory evaluations, vital signs and physical characteristics, safety in special groups and circumstances, including Adverse Events of Special Interest (AESI) (see Section 6.14.4), and study drug interruptions.

6.14.1. Adverse Events

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

A treatment-emergent adverse event (TEAE) is defined as an event that either first occurred or worsened in severity after the first dose of study treatment and on or prior to the date of the last visit within the treatment period being analyzed. Adverse events are classified based upon the MedDRA PT. The MedDRA Lowest Level Term (LLT) will be used in defining which events are treatment-emergent. The maximum severity for each LLT during the baseline period until the first dose of the study medication will be used as baseline. The treatment period (Parts A and B) will be included as postbaseline for the analysis. If an event is pre-existing during the baseline period but it has missing severity, and the event persists or reoccurs during the treatment period, then the baseline severity will be considered mild for determining any postbaseline treatment-emergence (ie, the event is treatment-emergent unless the severity is coded as mild postbaseline). If an event occurring postbaseline has a missing severity rating, then the event is considered treatment-emergent. Should there be insufficient data for AE start date to make this comparison (for example, the AE start year is the same as the treatment start year, but the AE start month and day are missing), the AE will be considered treatment-emergent. For events occurring on the day of the first dose of study treatment, a case report form (CRF) will be used to document whether the onset of the event was pre-treatment versus post-treatment in order to derive treatment-emergence.

For patients who were randomized to placebo and required rescue therapy or who were switched to baricitinib at the end of Part A, the AE severity just prior to rescue or switch to baricitinib will be used as baseline for assessing TEAE during baricitinib treatment.

Further, for some AEs of interest, duration of the TEAE may be summarized or listed within patient AE listings. The duration of each event will be defined as the end date of AE minus the start date of treatment-emergence of the AE plus 1 day. If an AE has not ended by the date of study drug discontinuation, early discontinuation of the study/period, or completion of the study, it will be censored as of that appropriate final date.

In general, summaries of TEAEs, SAEs, AEs that lead to discontinuation from the study, and AEs that lead to temporary interruption or permanent discontinuation of study drug will include the number of patients in the safety population (N), frequency of patients experiencing the event (n), and percent of patients experiencing the event (n/N x 100), and the exposure-adjusted incidence rate (EAIR). The EAIR will be expressed as the number of patients experiencing an

AE (n) per 100 patient years of exposure to treatment and is derived as 100 times the number of patients with the event divided by the sum of all patient exposure times (in years) to the treatment for the study/period. For any events that are gender-specific based on the displayed PT, the denominator used to compute the percentage will only include patients from the given gender. Treatment comparisons between treatment groups will be made using the Fisher exact test.

In an overview table, the number and percentage of patients in the safety population who experienced a death, a serious adverse event (SAE) (International Conference on Harmonisation [ICH]-defined, protocol-defined, and ICH- or protocol-defined), discontinuation from the study due to an AE or death, permanent discontinuation from study drug due to an AE or death, temporary interruption from study drug due to an AE, temporary interruption of study drug due to a lab abnormality, a TEAE, a TEAE rated as severe, or an AE possibly related to study drug based on the opinion of the investigator will be summarized by treatment.

The number and percent of patients with TEAEs will be summarized by treatment group using MedDRA PT nested within SOC. The SOCs will be ordered alphabetically, and events will be ordered within each SOC by decreasing frequency in the baricitinib group (or combined baricitinib group, as appropriate). As additional tables, the number and percent of patients with TEAEs will be summarized by treatment using MedDRA PT (without regard to SOC) and the number and percent of patients with TEAEs will be summarized by treatment using only the MedDRA SOC. Events will be ordered by decreasing frequency in the baricitinib group.

The number and percent of patients with TEAEs will be summarized by maximum severity by treatment using MedDRA PT nested within SOC for all TEAEs. For each patient and TEAE, the maximum severity for the MedDRA level being displayed (PT and SOC) is the maximum postbaseline severity observed from all associated LLTs mapping to that MedDRA level. For the "Severe" category, comparisons between treatment groups will be made using the Fisher exact test for each PT and SOC.

Adverse events considered possibly related to study drug by the investigator will be summarized by treatment group using MedDRA PT nested within SOC. A possibly-related event is defined as an event recorded on the eCRF as “Yes” to the question “Is the event possibly related to study drug?” Events will be ordered by decreasing frequency in the baricitinib group (or combined baricitinib group, where appropriate), within alphabetical SOC. If the investigator’s determination of possible relationship to study drug is missing, then the event will be considered as “related”.

Follow-up emergent adverse events (FEAEs) will be established for the post-treatment follow-up Period. An FEAE is an event that either first occurred or worsened in severity after the last visit during the treatment period. The baseline severity for FEAE is the severity level at the last visit from the treatment period; otherwise, FEAE processing is similar to TEAE processing. Follow-up AEs will be summarized in a similar manner as TEAEs (using MedDRA PT within SOC) according to groups defined by the treatment the patients were receiving prior to entering the follow-up period.

An individual listing of all AEs including preexisting conditions will be provided. A separate listing will include all AEs that led to discontinuation from the study.

6.14.1.1. Common Adverse Events

The number and percent of patients with TEAEs, including the EAIR, will be summarized by treatment group using MedDRA PT nested within SOC for common TEAEs (ie, TEAEs that occurred in ≥2% (before rounding) of any treatment group. Events will be ordered by decreasing frequency in the baricitinib group (or combined baricitinib group, where appropriate) within alphabetical SOC.

6.14.1.2. Serious Adverse Events

Consistent with the ICH E2A guideline, an SAE is any AE that results in 1 of the following outcomes:

- Death
- Initial or prolonged inpatient hospitalization
- A life-threating experience (ie, immediate risk of dying)
- Persistent or significant disability/incapacity
- Congenital anomaly/birth defect
- Considered significant by the investigator for any other reason.

In addition, if any patient experiences an AE that requires permanent discontinuation from study drug, for an AE or abnormal laboratory result, that AE or abnormal laboratory value will be reported as an SAE (reported as ‘Other reason serious’). These protocol-defined specific SAEs were defined as an AE or laboratory value that results in permanent study drug discontinuation or discontinuation from the study. These will be distinguished from ICH-defined SAEs using a scripted Clinical Data Question and Answer (CDQA) form to be completed by the investigator. The CDQA form will be sent as a query to the investigative site whenever an SAE is designated as serious based only on ‘other reason serious.’ The site will provide a response that will explain if the SAE was serious by conventional ICH criteria or if it was designated as serious only by the protocol requirement. In the absence of a CDQA response, the SAE cannot be considered as serious only by the protocol requirement and will be included in SAEs that are serious according to conventional ICH criteria.

The number and percent of patients who experienced any SAE (also ICH-defined and protocol-defined SAE) will be summarized by treatment group using MedDRA PT nested within SOC. Events will be ordered by decreasing frequency in the baricitinib group (or combined baricitinib group, where appropriate) within alphabetical SOC. Serious AEs overall and of each type will be summarized by treatment period and during the post-treatment follow-up period. An individual listing of all SAEs will be provided.

6.14.1.3. Other Significant Adverse Events

The number and percent of patients who discontinued from the study because of an AE or death, the number and percent of patients who permanently discontinued study drug because of an AE or death, and the number and percent of patients who temporarily interrupted study drug because of an AE will be summarized by treatment group using MedDRA PT nested within SOC. Events will be ordered by decreasing frequency in the baricitinib group (or combined baricitinib group, where appropriate) within alphabetical SOC.

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

6.14.2. Vital Signs and Physical Characteristics

Vital signs and physical characteristics include systolic blood pressure (SBP), diastolic blood pressure (DBP), pulse, weight, BMI, and waist circumference. Triplicate assessments of SBP and DBP are collected at each visit. When replicate assessments are recorded on the eCRF, the average will be derived; otherwise, the average which is recorded on the eCRF will be used for the summaries. Only the average across the replicates or the average recorded on the eCRF will be analyzed. When these parameters are analyzed as continuous variables at each visit, unscheduled measurements will be excluded. When these parameters are analyzed as categorical outcomes, as treatment-emergent abnormalities, and/or as continuous variables ignoring the visit (ie, change from minimum or maximum baseline value to minimum or maximum value during the treatment period), scheduled and unscheduled measurements will be included.

In general, vital signs and physical characteristics will be summarized with descriptive statistics (n, mean, SD, minimum, median, and maximum). Summaries will be provided for baseline and for each postbaseline visit by treatment group, including change from baseline and percent change from baseline. These summaries will be based on patients who have both baseline and at least 1 postbaseline assessment.

In addition, 3 approaches to the evaluation of change from baseline will be taken:

- Change from baseline to last observation during the treatment period considering baseline as the last non-missing observation from the baseline period;
- Change from baseline to the minimum value during the treatment period considering baseline as the minimum value of the non-missing observations in the baseline period;
- Change from baseline to the maximum value during the treatment considering baseline as the maximum value of the non-missing observations in the baseline period.

Change from baseline in vital signs and physical characteristics parameters, using each approach above, will be summarized using standard descriptive statistics by visit through Week 52.

The number and percent of patients with treatment-emergent high or low vital signs and weight at any time may be summarized by treatment group.

The treatment-emergent high or low vital signs and weight outcomes are:

- A treatment-emergent **high** result is defined as a change from a value less than or equal to the high limit at all baseline visits to a value greater than the high limit at any time that meets the specified change criteria in Table 6-2 during the treatment period.

- A treatment-emergent **low** result is defined as a change from a value greater than or equal to the low limit at all baseline visits to a value less than the low limit at any time that meets the specified change criteria in Table 6-2 during the treatment period.
- Treatment-emergent weight gain/loss is based on percent increase/decrease from baseline.

The number and percent of patients with treatment-emergent high or low vital signs may be further summarized in terms of shift tables by time point and last visit and by treatment group. Only scheduled measurements will be included in these analyses.

Table 6-2 defines the low and high baseline values as well as the criteria used to define treatment-emergence based on post-baseline values.

Table 6-2. Categorical Criteria for Abnormal Treatment-Emergent Blood Pressure and Heart Rate Measurement, and Categorical Criteria for Weight Changes for Adults

Parameter	Low	High
SBP (mm Hg) (sitting)	≤90 (low limit) and decrease from lowest value during baseline ≥20 if >90 at each baseline visit	≥140 (high limit) and increase from highest value during baseline ≥20 if <140 at each baseline visit
DBP (mm Hg) (sitting)	≤50(low limit) and decrease from lowest value during baseline ≥10 if >50 at each baseline visit	≥90 (high limit) and increase from highest value during baseline ≥10 if <90 at each baseline visit
Heart rate (bpm) (sitting)	<50 (low limit) and decrease from lowest value during baseline ≥15 if ≥50 at each baseline visit	>100 (high limit) and increase from highest value during baseline ≥15 if ≤100 at each baseline visit
Weight (kg)	(Loss) decrease ≥7% from lowest value during baseline	(Gain) increase ≥7% from highest value during baseline

Abbreviations: bpm = beats per minute; DBP = diastolic blood pressure; SBP = systolic blood pressure.

6.14.3. Laboratory Measurements

All laboratory tests will be presented using Système International of units (SI units). For topics of safety in special groups and circumstances, laboratory test units will be specified for each analysis.

Lilly large clinical trial population based reference limits (Lilly-defined; LCTPB) will be used to define the low and high limits since it is generally desirable for limits used for analyses to have greater specificity (identify fewer false positive cases) than reference limits used for individual patient management. For the hepatic laboratory assessments (ALT, AST, total bilirubin, and alkaline phosphatase [ALP]), central laboratory reference ranges (Quintiles) will be used and all results pertaining to these assessments will be included as a separate analysis to address the risk of liver injury as a special safety topic (see Section 6.14.4.1). Central laboratory reference ranges (Quintiles) will also be used to evaluate immunoglobulins (see Section 6.14.3).

The ratio of the low-density lipoprotein (LDL) cholesterol value to the high-density lipoprotein (HDL) cholesterol, referred to as the LDL/HDL ratio, will be derived for each patient at each visit where both values are simultaneously available. This lab parameter was not included in the standard lab processing for the Phase 3 RA studies and, as such, will not have valid reference ranges for comparison of the ratio to a standard for characterization as treatment-emergent abnormal.

Change from baseline to last observation for laboratory tests will be summarized for patients who have both baseline and at least 1 postbaseline result. Baseline will be the last nonmissing observation in the baseline period prior to receiving the first dose of the study drug. The last nonmissing observation in the treatment period will be used for this analysis. Unscheduled measurements will be excluded.

Change from the minimum value during the baseline period to the minimum value during the treatment period for selected laboratory tests will be summarized for patients who have both baseline and at least 1 post-baseline result. Baseline will be the minimum of non-missing observations in the baseline period. The minimum value in the treatment period will be used for this analysis. Similarly, change from the maximum value during the baseline period to the maximum value during the treatment period for selected laboratory tests will be summarized for patients who have both baseline and at least 1 post-baseline result. Baseline will be the maximum of non-missing observations in the baseline period. The maximum value in the treatment period will be used for this analysis. Original-scale data will be analyzed. Scheduled and unscheduled measurements will be included.

The change from baseline in laboratory assessments, using each of the 3 approaches above, will be summarized using standard descriptive statistics.

In addition, laboratory assessments will be summarized with descriptive statistics (n, mean, SD, minimum, median, and maximum) by time point. Summaries will be provided for baseline and for each postbaseline visit by treatment group, including change from baseline and percent change from baseline. These summaries will be based on patients who have both baseline and at least 1 postbaseline assessment.

The number and percent of patients with treatment-emergent abnormal, high, or low laboratory results at any time will be summarized by treatment group.

- A treatment-emergent **abnormal** result is defined as a change from normal at all baseline visits to abnormal at any time during the treatment period.
- A treatment-emergent **high** result is defined as a change from a value less than or equal to the high limit at all baseline visits to a value greater than the high limit at any time during the treatment period.
- A treatment-emergent **low** result is defined as a change from a value greater than or equal to the low limit at all baseline visits to a value less than the low limit at any time during the treatment period.

Scheduled and unscheduled measurements will be included in analysis of treatment-emergent abnormal, high, or low laboratory results. Differences between treatments (when appropriate) in incidence of treatment-emergent abnormal, high, or low laboratory results will be assessed using the Fisher exact test.

Change from study baseline to final treatment visit, from final treatment visit to follow-up visit, and from study baseline to follow-up visit in laboratory assessments will be summarized using standard descriptive statistics. The number and percent of patients with treatment-emergent high, normal, or low laboratory results will be further summarized comparing study baseline to last follow-up visit and follow-up baseline to last follow-up visit by treatment group.

Individual listings of laboratory results which include hematology, chemistry, urinalysis, and immunoglobulin will be provided. Additional listings that restrict to abnormal results will also be provided for both the treatment period and the post-treatment follow-up period. All lab listings will be performed using SI units.

6.14.4. Safety in Special Groups and Circumstances, including Adverse Events of Special Interest

In addition to general safety parameters, safety information on specific topics of special interest will be presented. Topics will be identified by laboratory results, standardized MedDRA queries (SMQs), Lilly-defined MedDRA PT listings, or a combination of these methods. The primary focus of analyses will be on laboratory test values, where applicable, with a secondary focus on reported AEs related to laboratory values. Additional special safety topics may be added as warranted. The topics outlined in this section include the protocol-specified AESI:

- severe or potential opportunistic infections
- myelosuppressive events of anemia, leukopenia, neutropenia, lymphopenia, and thrombocytopenia
- thrombocytosis
- elevations in ALT or AST (>3 × ULN) with total bilirubin (>2 × ULN).

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

6.14.4.1. Abnormal Hepatic Tests

Analyses for abnormal hepatic tests involve 5 lab analytes: ALT, AST, total bilirubin, ALP, and albumin. Analyses for change from baseline to last observation, change from the minimum value during the baseline period to the minimum value during the treatment period, change from the maximum value during the baseline period to the maximum value during the treatment period, and treatment-emergent high or low laboratory results at any time are described in Section 6.14.3. This section describes additional analyses for the topic. Recall also that central laboratory reference ranges (Quintiles) will be used for some hepatic laboratory assessments (ALT, AST, total bilirubin, and ALP).

The number and percent of patients with the following elevations in hepatic laboratory tests at any time will be summarized between treatment groups:

- The percentages of patients with an ALT measurement ≥3 ×, 5 ×, and 10 × the central laboratory ULN during the treatment period will be summarized for all patients with a postbaseline value and for subsets based on various levels of baseline.

  - The analysis of 3 × ULN will contain 4 subsets: patients whose nonmissing maximum baseline value is ≤1 × ULN, patients whose maximum baseline is >1 × ULN but <3 × ULN, patients whose maximum baseline value is ≥3 × ULN, and patients whose baseline values are missing.
  - The analysis of 5 × ULN will contain 5 subsets: patients whose nonmissing maximum baseline value is ≤1 × ULN, patients whose maximum baseline is >1 × ULN but <3 × ULN, patients whose maximum baseline is ≥3 × ULN but <5 × ULN, patients whose maximum baseline value is ≥5 × ULN, and patients whose baseline values are missing.
  - The analysis of 10 × ULN will contain 6 subsets: patients whose nonmissing maximum baseline value is ≤1 × ULN, patients whose maximum baseline is >1 × ULN but <3 × ULN, patients whose maximum baseline is ≥3 × ULN but <5 × ULN, patients whose maximum baseline is ≥5 × ULN but <10 × ULN, patients whose maximum baseline value is ≥to 10 × ULN, and patients whose baseline values are missing.

- The percentages of patients with an AST measurement ≥3 ×, 5 ×, and 10 × the central laboratory ULN during the treatment period will be summarized for all patients with a postbaseline value and for subsets based on various levels of baseline. Analyses will be constructed as described above for ALT.
- The percentages of patients with a total bilirubin measurement ≥2 × the central laboratory ULN during the treatment period will be summarized for all patients with a postbaseline value and subset into 4 subsets: patients whose nonmissing maximum baseline value is ≤1 × ULN, patients whose maximum baseline is >1 × ULN but <2 × ULN, patients whose maximum baseline value is ≥2 × ULN, and patients whose baseline values are missing.
- The percentages of patients with an ALP measurement ≥1.5 × the central laboratory ULN during the treatment period will be summarized for all patients with a postbaseline value and subset into 4 subsets: patients whose nonmissing maximum baseline value is ≤1 × ULN, patients whose maximum baseline is >1× ULN but <1.5 × ULN, patients whose maximum baseline value is ≥1.5 × ULN, and patients whose baseline values are missing.

A by-patient listing will be provided for an ALT or AST measurement ≥3 × ULN, or with an ALP measurement ≥1.5 × ULN, or with a total bilirubin measurement ≥2 × ULN. The listing will include relevant demographic, study drug, concomitant therapy, AEs, and laboratory data.

To further evaluate potential hepatotoxicity, an Evaluation of Drug-Induced Serious Hepatotoxicity (eDISH) plot will be created. Each patient with at least 1 post-baseline ALT and total bilirubin contributes 1 point to the plot. The points correspond to maximum total bilirubin and maximum ALT, even if not obtained from the same blood draw.

Criteria for AE grading from the Common Terminology Criteria for Adverse Events (CTCAE) will be applied for the hepatic laboratory tests AST, ALT, ALP, total bilirubin, and albumin (Table 6-3). These CTCAE grading schemes are consistent with both Version 3.0 and Version

4.03 of the CTCAE guidelines. When deriving CTCAE grades for laboratory analytes with absolute criteria for grading, such as albumin, an adjustment to the CTCAE-defined approach will be made. The use of a lower limit of normal/upper limit of normal (LLN/ULN) from appropriate reference ranges will be replaced with a simpler approach that uses a single value. This adjustment to the CTCAE grading criteria overcomes the irresolute derivation of the grade that can occur when some demographically-specific LLN/ULN values (from either the Lilly LCTPB or the Quintiles reference ranges) overlap with the CTCAE-defined value that separates normal (Grade 0) and Grade 1.

Table 6-3. Common Terminology Criteria for Adverse Events Related to Hepatic Tests

Laboratory Test	Laboratory Units	Grade	Criteria
Albumin	SI	0 (normal)	≥35 g/L
		1	<35 g/L and ≥30 g/L
		2	<30 g/L and ≥20 g/L
		3	<20 g/L
ALT	SI	0 (normal)	≤ULN
		1	>ULN and ≤2.5× ULN
		2	>2.5× ULN and ≤5× ULN
		3	>5× ULN and ≤20× ULN
		4	>20× ULN
AST	SI	0 (normal)	≤ULN
		1	>ULN and ≤2.5× ULN
		2	>2.5× ULN and ≤5× ULN
		3	>5× ULN and ≤20× ULN
		4	>20× ULN
ALP	SI	0 (normal)	≤ULN
		1	>ULN and ≤.5× ULN
		2	>2.5× ULN and ≤5× ULN
		3	>5× ULN and ≤20× ULN
		4	>20× ULN
Total bilirubin	SI	0 (normal)	≤ULN
		1	>ULN and ≤1.5× ULN
		2	>1.5× ULN and ≤3× ULN
		3	>3× ULN and ≤10× ULN
		4	>10× ULN

Abbreviations: ALP = alkaline phosphatase; ALT = alanine transaminase; AST = aspartate aminotransferase; ULN = upper limit of normal.

Treatment-emergent laboratory abnormalities related to hepatic abnormalities occurring at any time during the treatment period will be tabulated. Treatment-emergence will be characterized using 2 criteria (as appropriate to the grading scheme):

- any increase in post-baseline CTCAE grade from worst baseline grade
- an increase from worst baseline less than Grade 3 to Grade 3 or Grade 4 at worst post-baseline

Treatment comparisons will be made using the Fisher exact test as appropriate. Scheduled and unscheduled measurements will be included.

Treatment-emergent laboratory abnormalities related to hepatic abnormalities shift tables of baseline to maximum during the treatment period will be tabulated. Shift tables will show the number and percent of patients with measurement in each category based on baseline to maximum CTCAE grade during the treatment period (Part A and B), with baseline depicted by the most extreme grade during the baseline period. A shift table summary displaying the number and percent of patients with maximum postbaseline results will be presented overall and by treatment group within the following categories:

- Decreased; post-baseline category < baseline category,
- Increased; post-baseline category > baseline category,
- Same; post-baseline category = baseline category.

The number and percent of patients with treatment-emergent hepatic laboratory abnormalities based on CTCAE grades will be summarized in terms of shift tables descriptively comparing baseline to maximum and last post-treatment follow-up grade and follow-up baseline to maximum and last post-treatment follow-up grade by treatment group. A shift table summary for each analysis will also be included. Furthermore, a summary of grade increases by each maximum and last grade will be provided.

First shift table summary:

- Decreased; maximum post-treatment follow-up category < baseline category
- Increased; maximum post-treatment follow-up category > baseline category
- Same; maximum post-treatment follow-up category = baseline category

Second shift table summary:

- Decreased; maximum post-treatment follow-up category < follow-up baseline category
- Increased; maximum post-treatment follow-up category > follow-up baseline category
- Same; maximum post-treatment follow-up category = follow-up baseline category

Third shift table summary:

- Decreased; last post-treatment follow-up category < baseline category
- Increased; last post-treatment follow-up category > baseline category
- Same; last post-treatment follow-up category = baseline category

Fourth shift table summary:

- Decreased; last post-treatment follow-up category < follow-up baseline category
- Increased; last post-treatment follow-up category > follow-up baseline category
- Same; last post-treatment follow-up category = follow-up baseline category

6.14.4.2. Myelosuppressive Events

Using clinical laboratory assessments, the myelosuppressive events of anemia, leukopenia, neutropenia, lymphopenia, and thrombocytopenia will be assessed through analysis of decreased hemoglobin, decreased white cell count, decreased neutrophils absolute count, decreased lymphocytes absolute count, and decreased platelet count, respectively.

Laboratory tests for these measures may be summarized across time points using boxplots which include accompanying summary statistics.

Criteria for AE grading from the CTCAEs will be applied for laboratory tests related to myelosuppressive events (Table 6-4). These CTCAE grading schemes are consistent with both Version 3.0 and Version 4.03 of the CTCAE guidelines. When deriving CTCAE grades for laboratory analytes with absolute criteria for grading, an adjustment to the CTCAE-defined approach will be made. The use of an LLN/ULN from appropriate reference ranges will be replaced with a simpler approach that uses a single value (or a sex-specific pair of values for hemoglobin). This adjustment to the CTCAE grading criteria overcomes the irresolute derivation of the grade that can occur when some demographically-specific LLN/ULN values overlap with the CTCAE-defined value that separates normal (Grade 0) and Grade 1.

Treatment-emergence will be characterized using 2 criteria:

- any increase in post-baseline CTCAE grade from worst baseline grade
- an increase from worst baseline less than Grade 3 to Grade 3 or Grade 4 at worst post-baseline

Table 6-4. Common Terminology Criteria for Adverse Events Related to Myelosuppressive Events

Event	Laboratory Test	Grade	Criteria in SI Units
Anemia	Hemoglobin	0 (normal)	≥7.27 mmol (Fe)/L for females and ≥8.18 mmol (Fe)/L for males
		1	<7.27 mmol (Fe)/L for females and <8.18 mmol (Fe)/L for males and ≥6.2 mmol (Fe)/L for both females and males
		2	<6.2 mmol (Fe)/L and ≥4.9 mmol (Fe)/L
		3	<4.9 mmol (Fe)/L and ≥4.0 mmol (Fe)/L
		4	<4.0 mmol (Fe)/L
Leukopenia	White blood cell (WBC) count	0 (normal)	≥4.0 GI/L
		1	<4.0 GI/L and ≥3.0 GI/L
		2	<3.0 GI/L and ≥2.0 GI/L
		3	<2.0 GI/L and ≥1.0 GI/L
		4	<1.0 GI/L
Neutropenia	Absolute neutrophil count (ANC)	0 (normal)	≥2 GI/L
		1	<2 GI/L and ≥1.5 GI/L
		2	<1.5 GI/L and ≥1.0 GI/L
		3	<1.0 GI/L and ≥0.5 GI/L
		4	<0.5 GI/L
Lymphopenia	Lymphocyte count	0 (normal)	≥1.1 GI/L
		1	<1.1 GI/L and ≥0.8 GI/L
		2	<0.8 GI/L and ≥0.5 GI/L
		3	<0.5 GI/L and ≥0.2 GI/L
		4	<0.2 GI/L
Thrombocytopenia	Platelet count	0 (normal)	≥150 GI/L

	1	<150 GI/L and ≥75 GI/L
	2	<75 GI/L and ≥50 GI/L
	3	<50 GI/L and ≥25 GI/L
	4	<25 GI/L

Abbreviations: ANC = absolute neutrophil count; dL = deciliters; Fe = iron; g = grams; GI/L = Giga/Liter; mmol = millimoles; SI = Système International; WBC = white blood cell.

A laboratory-based treatment emergent outcome related to increased platelet count will be summarized in similar fashion. Treatment-emergent thrombocytosis as a laboratory-based abnormality will be defined as an increase in platelet count from a maximum baseline value ≤600,000/μL to any postbaseline value >600,000/μL. A frequency table showing the number and percent of patients meeting criteria for thrombocytosis by treatment group will be provided. Scheduled and unscheduled measurements will be included.

Shift tables will show the number and percent of patients with measurement in each category based on baseline to maximum grade during the treatment period, with baseline depicted by the most extreme grade during the baseline period. With each shift table, a shift table summary will display the number and percent of patients with maximum postbaseline results overall and by treatment group within the following categories:

- Decreased; postbaseline category < baseline category,
- Increased; postbaseline category > baseline category,
- Same; postbaseline category = baseline category.

Scheduled and unscheduled measurements will be included. Treatment differences for the shift table summary categorizations may be assessed with the Likelihood Ratio Chi-Square test, using the exact version of this test whenever at least 1 cell has an expected cell count <5. Furthermore, a summary of grade increases by each maximum grade will be provided.

The number and percent of patients with measurement in each category based on baseline to CTCAE grade at each scheduled visit during the treatment period, with baseline depicted by the last grade during the baseline period, will be further summarized in terms of shift tables by time point and last visit and by treatment group. Only scheduled measurements will be included in these analyses. Following the shift table for the last visit, a shift table summary of the last visit will be included. Furthermore, a summary of grade increases by each grade at last visit will be provided.

The number and percent of patients with treatment-emergent lab abnormalities related to myelosuppression based on CTCAE grades will be summarized in terms of shift tables descriptively comparing baseline to maximum and last post-treatment follow-up grade and follow-up baseline to maximum and last post-treatment follow-up grade by treatment group. A shift table summary for each analysis will also be included. Furthermore, a summary of grade increases by each maximum and last grade will be provided.

First shift table summary:

- Decreased; maximum post-treatment follow-up category < baseline category

- Increased; maximum post-treatment follow-up category > baseline category
- Same; maximum post-treatment follow-up category = baseline category

Second shift table summary:

- Decreased; maximum post-treatment follow-up category < follow-up baseline category
- Increased; maximum post-treatment follow-up category > follow-up baseline category
- Same; maximum post-treatment follow-up category = follow-up baseline category

Third shift table summary:

- Decreased; last post-treatment follow-up category < baseline category
- Increased; last post-treatment follow-up category > baseline category
- Same; last post-treatment follow-up category = baseline category

Fourth shift table summary:

- Decreased; last post-treatment follow-up category < follow-up baseline category
- Increased; last post-treatment follow-up category > follow-up baseline category
- Same; last post-treatment follow-up category = follow-up baseline category

6.14.4.3. Lipids Effects

Lipids effects will be assessed through analysis of elevated total cholesterol, elevated LDL cholesterol, decreased HDL cholesterol, and elevated triglycerides.

National Cholesterol Education Program (NCEP) Adult Treatment Panel III will be applied for laboratory tests related to lipids effects (Table 6-5). The grade-like categories shown in this table are ordered from lowest to highest for the purposes of these analyses.

Table 6-5. National Cholesterol Education Adult Treatment Panel III Criteria

Lab Test	Category	Criteria in SI Units
Total Cholesterol	Desirable	<5.17 mmol/L
	Borderline high	≥5.17 mmol/L and <6.21 mmol/L
	High	≥6.21 mmol/L
LDL–C	Optimal	<2.59 mmol/L
	Near optimal	≥2.59 mmol/L and <3.36 mmol/L
	Borderline high	≥3.36 mmol/L and <4.14 mmol/L
	High	≥4.14 mmol/L and <4.91 mmol/L
	Very high	≥4.91 mmol/L
HDL-C	High	≥1.55 mmol/L
	Normal	≥1.03 mmol/L and <1.55 mmol/L
	Low	<1.03 mmol/L
Triglycerides	Normal	<1.69 mmol/L
	Borderline high	≥1.69 mmol/L and <2.26 mmol/L
	High	≥2.26 mmol/L and <5.65 mmol/L
	Very high	≥5.65 mmol/L

Abbreviations: HDL-C = high-density lipoprotein cholesterol; LDL-C = low-density lipoprotein cholesterol.

Treatment-emergent laboratory abnormalities related to elevated total cholesterol, elevated LDL cholesterol, decreased HDL cholesterol, and elevated triglycerides occurring at any time during the treatment period will be tabulated using the NCEP categories shown in Table 6-5. Treatment-emergent elevated total cholesterol will be characterized as movement from categories ‘Desirable’ or ‘Borderline High’ to category ‘High’ considering patients who are at risk for the abnormality, that is, not ‘High’ based on the maximum baseline total cholesterol category. Treatment-emergent elevated LDL cholesterol will be characterized as movement from categories ‘Optimal,’ ‘Near optimal,’ or ‘Borderline high’ to categories ‘High’ or ‘Very high’ considering patients who are at risk for the abnormality, that is, neither ‘High’ nor ‘Very high’ based on the maximum baseline LDL cholesterol category. Treatment-emergent decreased HDL cholesterol will be characterized as movement from categories ‘High’ or ‘Normal’ to category ‘Low’ considering patients who are at risk for the abnormality, that is, not ‘Low’ based on the minimum baseline HDL cholesterol category. Treatment-emergent elevated triglycerides will be characterized as movement from categories ‘Normal’ or ‘Borderline high’ to categories ‘High’ or ‘Very high’ considering patients who are at risk for the abnormality, that is, neither ‘High’ nor ‘Very high’ based on the maximum baseline triglycerides category. Scheduled and unscheduled measurements will be included.

Shift tables will show the number and percent of patients with measurement in each category based on baseline to maximum or minimum category during the treatment period, with baseline depicted by the most extreme category during the baseline period. With each shift table, a shift table summary will display the number and percentage of patients with maximum or minimum postbaseline results overall and by treatment group within the following categories:

- Decreased; postbaseline category < baseline category,
- Increased; postbaseline category > baseline category,
- Same; postbaseline category = baseline category.

Treatment differences for the shift table summary categorizations may be assessed with the Likelihood Ratio Chi-Square test, using the exact version of this test whenever at least 1 cell has an expected cell count <5. Furthermore, a summary of category increases by each maximum category will be provided.

In addition, the number and percent of patients with measurement in each category based on baseline to NCEP category at each scheduled time point, with baseline depicted by the last category during the baseline period, will be further summarized in terms of shift tables by time point and last visit and by treatment group. Only scheduled measurements will be included in these analyses. Following the shift table for the last visit, a shift table summary of the last visit will be included. Furthermore, a summary of category increases by each category at last visit will be provided.

6.14.4.4. Renal Function Effects

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

Laboratory tests for creatinine may be summarized across time points using boxplots which include accompanying summary statistics. Unscheduled measurements will be excluded from these analyses.

Common Terminology Criteria for Adverse Events will be applied for laboratory tests related to renal effects (Table 6-6). The creatinine grades come from CTCAE Version 3.0 guidelines.

Treatment-emergent laboratory abnormalities related to elevated creatinine occurring at any time during the treatment period will be tabulated using the CTCAE grades shown in Table 6-6. Scheduled and unscheduled measurements will be included.

Treatment-emergence will be characterized using 2 characterizations:

- any increase in postbaseline CTCAE grade from baseline grade
- an increase from baseline less than Grade 3 to Grade 3 or Grade 4 at worst postbaseline

Table 6-6. Common Terminology Criteria for Adverse Events Related to Renal Effects

Lab Test	CTCAE Version	Grade	Criteria
Elevated creatinine	3.0	0 (nomal)	≤ULN
		1	>ULN and ≤1.5× ULN
		2	>1.5× ULN and ≤3× ULN
		3	>3× ULN and ≤6× ULN
		4	>6× ULN

Abbreviations: ULN = upper limit of normal.

The number and percent of patients with measurement in each category based on baseline to CTCAE grade at each scheduled time point, with baseline depicted by the last grade during the baseline period, will be further summarized in terms of shift tables by time point and last visit and by treatment group. Only scheduled measurements will be included in these analyses. Following the shift table for the last visit, a shift table summary of the last visit will be included. Furthermore, a summary of grade increases by each grade at last visit will be provided.

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

- Decreased; postbaseline category < baseline category,
- Increased; postbaseline category > baseline category,
- Same; postbaseline category = baseline category.

Treatment differences for the shift table summary categorizations will be assessed with the Likelihood Ratio Chi-Square test, using the exact version of this test whenever at least 1 cell has an expected cell count <5. Scheduled and unscheduled measurements will be included. Furthermore, a summary of grade increases by each maximum grade will be provided.

The number and percent of patients with treatment-emergent lab abnormalities related to elevated creatinine based on CTCAE grades will be summarized in terms of shift tables descriptively comparing baseline to maximum and last post-treatment follow-up and follow-up baseline to maximum and last post-treatment follow-up by treatment group. A shift table summary for each analysis will also be included. Furthermore, a summary of grade increases by each maximum and last grade will be provided.

First shift table summary:

- Decreased; maximum post-treatment follow-up category < baseline category
- Increased; maximum post-treatment follow-up category > baseline category
- Same; maximum post-treatment follow-up category = baseline category

Second shift table summary:

- Decreased; maximum post-treatment follow-up category < follow-up baseline category
- Increased; maximum post-treatment follow-up category > follow-up baseline category
- Same; maximum post-treatment follow-up category = follow-up baseline category

Third shift table summary:

- Decreased; last post-treatment follow-up category < baseline category
- Increased; last post-treatment follow-up category > baseline category
- Same; last post-treatment follow-up category = baseline category

Fourth shift table summary:

- Decreased; last post-treatment follow-up category < follow-up baseline category
- Increased; last post-treatment follow-up category > follow-up baseline category
- Same; last post-treatment follow-up category = follow-up baseline category

6.14.4.5. Elevations in Creatine Phosphokinase

Laboratory tests for creatine phosphokinase (CPK) may be summarized across time points using boxplots which include accompanying summary statistics. Unscheduled measurements will be excluded from these analyses.

Elevations in CPK will be addressed using CTCAE criteria (Table 6-7) from CTCAE Version 3.0 guidelines.

Treatment-emergent laboratory abnormalities related to elevated CPK occurring at any time during the treatment period will be tabulated using the CTCAE grades shown in Table 6-7.

Treatment-emergence will be characterized using 2 characterizations:

- any increase in postbaseline CTCAE grade from worst baseline grade
- an increase from worst baseline less than Grade 3 to Grade 3 or Grade 4 at worst postbaseline

Scheduled and unscheduled measurements will be included.

Table 6-7. Common Terminology Criteria for Adverse Events Related to Elevations in Creatine Phosphokinase

Lab Test	CTCAE Version	Grade	Criteria
Elevated CPK	3.0	0 (normal)	≤ULN
		1	>ULN and ≤2.5× ULN
		2	>2.5× ULN and ≤5× ULN
		3	5× ULN and ≤10× ULN
		4	>10× ULN

Abbreviations: CPK = creatine phosphokinase; LCTPB = large clinical trial population based; ULN = upper limit of normal based on Lilly LCTPB reference limits.

Shift tables will show the number and percent of patients with measurement in each category based on baseline to maximum CTCAE grade during the treatment period, with baseline depicted by the most extreme grade during the baseline period. With each shift table, a shift table summary will display the number and percent of patients with maximum postbaseline results, by treatment group within the following categories:

- Decreased; postbaseline category < baseline category
- Increased; postbaseline category > baseline category
- Same; postbaseline category = baseline category.

Treatment differences for the shift table summary categorizations will be assessed with the Likelihood Ratio Chi-Square test, using the exact version of this test whenever at least 1 cell has an expected cell count <5. Scheduled and unscheduled measurements will be included. Furthermore, a summary of grade increases by each maximum grade will be provided.

In addition, the number and percent of patients with measurement in each category based on baseline to CTCAE grade at each scheduled visit, with baseline depicted by the last grade during the baseline period, will be further summarized in terms of shift tables by time point and last visit and by treatment group. Only scheduled measurements will be included in these analyses. Following the shift table for the last visit, a shift table summary of the last visit will be included. Furthermore, a summary of grade increases by each grade at last visit will be provided.

The number and percent of patients with treatment-emergent laboratory abnormalities related to elevated CPK based on CTCAE grades will be summarized in terms of shift tables descriptively comparing baseline to maximum and last post-treatment follow-up and follow-up baseline to maximum and last post-treatment follow-up by treatment group. A shift table summary for each analysis will also be included. Furthermore, a summary of grade increases by each maximum and last grade will be provided.

First shift table summary:

- Decreased; maximum post-treatment follow-up category < baseline category
- Increased; maximum post-treatment follow-up category > baseline category
- Same; maximum post-treatment follow-up category = baseline category

Second shift table summary:

- Decreased; maximum post-treatment follow-up category < follow-up baseline category
- Increased; maximum post-treatment follow-up category > follow-up baseline category
- Same; maximum post-treatment follow-up category = follow-up baseline category

Third shift table summary:

- Decreased; last post-treatment follow-up category < baseline category
- Increased; last post-treatment follow-up category > baseline category
- Same; last post-treatment follow-up category = baseline category

Fourth shift table summary:

- Decreased; last post-treatment follow-up category < follow-up baseline category
- Increased; last post-treatment follow-up category > follow-up baseline category
- Same; last post-treatment follow-up category = follow-up baseline category

6.14.4.6. Infections, Including Potential Opportunistic Infections

Infections will be defined using the PTs from the MedDRA Infections and Infestations SOC, with additional terms from the Investigations SOC being used in selected instances, as described below.

Treatment-emergent infections will be analyzed according to various groups of infectious events including:

- all infections
    - all PTs in the Infections and Infestations SOC,
- serious infections
    - all PTs in the Infections and Infestations SOC that are SAEs,
- infections that require therapeutic intervention (antibiotics, antivirals, antifungals, etc.)
    - all PTs in the Infections and Infestations SOC for which there is an antimicrobial concomitant medication associated with that event for that patient,
- herpes zoster
    - specific Lilly-defined PTs from the Herpes Viral Infections high-level term (HLT) in the Infections and Infestations SOC, shown in the appendices of the Baricitinib Program Safety Analysis Plan (PSAP), Version 4 (or most current version)
- tuberculosis
    - specific Lilly-defined PTs from the Tuberculous Infections HLT and the Atypical Mycobacterial infections HLT of the Infections and Infestations SOC and the Investigations SOC, shown in the appendices of the current version of the Baricitinib PSAP, Version 4 (or most current version)
- viral hepatitis
    - all PTs from the Hepatitis Viral Infections HLT (HLT code 10057212) in the Infections and Infestations SOC,
- potential opportunistic infections (POIs), as described in detail below.

For each of all infections, serious infections (overall and on each approach to identifying SAEs), infections that require therapeutic intervention, herpes zoster infections, tuberculosis, viral hepatitis, and potential opportunistic infections, the frequency and relative frequency for each PT will be provided, ordered by decreasing frequency in the baricitinib group (or combined baricitinib group, where appropriate).

In addition to the incidence of infectious AEs by MedDRA PT as described above, the number and percent of patients with treatment-emergent infectious AEs by infecting organism by treatment group will be summarized. Infecting organism is defined as any of the primary, secondary, or tertiary infecting organisms as entered on the eCRF. In addition, the frequency counts and percents of patients experiencing treatment-emergent infectious AEs will also be presented by treatment group and primary site of infection. The site of infection used in summary displays is based on a grouping of sites from the eCRF that is more meaningful than the many specific sites allowed on the eCRF. The PSAP shows how the eCRF-available entries are bundled into site categories for some reporting purposes. Summaries that display the infecting site within PT will use the actual site entered into the eCRF.

Potential Opportunistic Infections:

Potential Opportunistic Infections will be identified according to 2 different approaches.

First, POIs are identified from TEAEs based on a Lilly-defined list of MedDRA PTs shown in the PSAP, Version 4. These PTs are a subset of terms from the Infections and Infestations SOC. Note that the current list is based on MedDRA Version 18, but it will be updated for use as each MedDRA version is released, including MedDRA versions released after SAP finalization but before database lock.

Second, POIs are identified using a Lilly-defined list of combinations of the infectious organism and site of infection, as recorded on the specialized eCRF designed to provide additional details about infections; this list is contained in the PSAP, Version 4. Up to 3 infecting organisms are collected (primary, second, and third), and any of them can contribute to a POI. Only a single (primary) site is collected.

For final analysis, the most current version of PSAP will be used, including PSAP released after SAP finalization but before database lock.

For the POIs identified from MedDRA PTs, the number and percent of patients overall and for each specific PT will be summarized by treatment group, with specific event terms ordered by decreasing frequency in the baricitinib group (or combined baricitinib group, where appropriate).

For the POI identified from organism/site combinations, the number and percent of patients overall and for each specific combination will be summarized by treatment group, with specific organism/site pairs ordered by decreasing frequency of organism in the baricitinib group (or combined baricitinib group, where appropriate) and then alphabetically for sites within organism.

Treatment-emergent infectious events will be reviewed in context of other clinical and laboratory parameters. A listing of patients experiencing treatment-emergent infectious AEs will be provided, which will include patient demographics, treatment group, treatment start and stop dates, infectious event, event start and stop dates, event severity and possible relationship to the study drug, total leukocytes, total lymphocytes, absolute neutrophils, event seriousness, event outcome, whether the patient was immunized for tuberculosis and/or herpes zoster, and whether anti-microbial treatment was recorded.

6.14.4.7. Major Adverse Cardiovascular Events

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

Categories and subcategories analyzed will include the following:

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

In general, events requiring adjudication are documented by investigative sites using an endpoint reporting CRF. This CRF is then sent to the adjudication center which uses a cardiovascular adjudication reporting CRF to document the final assessment of the event as a MACE, as some other cardiovascular event, or as no event (according to the End point Source Document Submission Site Guide). In some cases, however, the investigator may not have deemed that an event had met the endpoint criteria but the event was still sent for adjudication as a potential MACE, other cardiovascular event, or no event. In these instances, the adjudication reporting CRF will not have a matching endpoint reporting CRF from the investigator. Events generated from these circumstances will be considered as events sent for adjudication in the absence of an investigator's endpoint reporting form.

First, the number and percent of patients with potential MACE and other potential cardiovascular events as sent for adjudication will be summarized by treatment group based on the categories and subcategories above. If an event is adjudicated without a matching endpoint reporting CRF from the investigative site, then these events will be separately summarized using the MedDRA PT for the reported event as documented on the adjudication reporting CRF. This term will be used in lieu of the term that would have been included on an endpoint data collection form. Treatment comparisons will be made using the Fisher exact test.

Second, the number and percent 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 above. Treatment comparisons may be made using the Fisher exact test.

Third, an overall summary of the number of events and deaths sent for adjudication by adjudicated outcome may be produced. The count of events (including death) sent for adjudication will be summarized along with how many of these were adjudicated as MACE, other cardiovascular event, or no event. A separate summary will be displayed for the numbers of deaths sent for adjudication and whether they were adjudicated as cardiovascular or non-cardiovascular.

A listing of all MACE or other cardiovascular events sent for adjudication will be provided and will include data concerning the reason the event was sent for adjudication, the adjudicated outcome, and the related MedDRA PT.

6.14.4.8. Malignancies

Malignancies will be identified using terms from the malignant tumours SMQ (SMQ 20000194).

A by-patient listing for all patients having a malignancy event at any time will be provided and will include the information whether or not the event is considered as SAE, its relatedness to study drug, and whether or not it led to discontinuation from the study, led to permanent study drug discontinuation, or led to temporary study drug interruption.

If the number of events in the study warrants a table summary, the number and percent of patients with TEAEs associated with malignant tumors will be summarized by treatment group using MedDRA PT. Events will be ordered by decreasing frequency in the baricitinib group (or combined baricitinib group, where appropriate). Additional summaries will give an overview of malignancies. First, the total number of patients with malignant tumors which are SAEs by ICH GCP or by protocol, by ICH GCP, and by protocol only, will be provided. Second, the number of patients with TEAEs associated with malignant tumors that were rated as severe, and that were considered related to study drug in the opinion of the investigator, will be provided. Third, the number of patients with TEAEs associated with malignant tumors that led to discontinuation from the study, that led to permanent study drug discontinuation, and that led to temporary study drug interruption will be provided. EAIR will also be calculated for each PT. Tumors that are gender-specific, based on PT, will have the denominators limited to patients of that gender. Treatment comparisons may be made using the Fisher exact test.

6.14.4.9. Allergic Reactions/Hypersensitivities

Allergic reactions/hypersensitivities will be examined within the context of anaphylactic reactions. Patients with a treatment-emergent anaphylactic reaction will be identified using the MedDRA Anaphylactic Reaction SMQ (SMQ 20000021). This SMQ is unique compared to most SMQs in that an algorithmic approach is taken to identify events.

Anaphylaxis has been broadly defined as “a serious allergic reaction that is rapid in onset and may cause death” (Sampson et al. 2006). Sampson described a two-tiered approach to identify cases of anaphylactic reaction, an approach which is mimicked by the algorithmic approach taken for the Anaphylactic Reaction SMQ.

The algorithm is defined in the following manner within the Introductory Guide for SMQs Version 17.0: “All narrow search terms as well as: if 1 term from category B and 1 term from category C is present, or if 1 term from (category B or category C) plus 1 term from category D.” This algorithm is also expressed as “A or (B and C) or (D and (B or C)).” The narrow search terms represent core anaphylactic reaction terms whereas Categories B, C, and D represent signs and symptoms that are possibly indicative of anaphylactic reaction, with Category B terms representing Upper Airway/Respiratory signs and symptoms, Category C representing Angioedema/Urticaria/Prutitis/Flush signs and symptoms, and Category D representing Cardiovascular/Hypertension signs and symptoms. Note that all terms within Categories B, C, and D are broad terms. The use of the narrow search terms match roughly with Sampson’s first tier approach, whereas the broad terms of Categories B, C, and D match roughly with Sampson’s second tier approach. Within this second approach using broad terms, it is important to recognize that occurrence of these events should be nearly coincident and develop rapidly after exposure to an antigen; based on recording of events on CRFs, a window wherein the events occur within 2 days of one another (based on the onset dates of the events) is allowed.

A by-patient listing for all patients having an allergic reaction/hypersensitivity event at any time will be provided and will include the information of event type, whether or not the event is considered as SAE, its relatedness to study drug, and whether or not it led to discontinuation from the study, led to permanent study drug discontinuation, or led to temporary study drug interruption.

If the number of events in the study warrants a table summary, the number and percent of patients with treatment-emergent anaphylactic reactions will be summarized by treatment group. Results for each PT will be provided along with subtotals that first pool narrow and broad terms together, and then, second, for narrow terms only, ordered by decreasing frequency in the baricitinib group (or combined baricitinib group, where appropriate). Additional summaries will give an overview of anaphylactic reactions. First, the total number of patients with anaphylactic reactions which are SAEs by ICH GCP or by protocol, by ICH GCP, and by protocol only, will be provided. Second, the number of patients with TEAEs associated with anaphylactic reactions that were rated as severe, and that were considered related to study drug in the opinion of the investigator, will be provided. Third, the number of patients with TEAEs associated with anaphylactic reactions that led to discontinuation from the study, that led to permanent study drug discontinuation, and that led to temporary study drug interruption will be provided.

6.14.4.10. Gastrointestinal Perforations

Treatment-emergent adverse events potentially related to gastrointestinal (GI) perforations will be analyzed using reported AEs. Identification of these events will be based on the PTs of the MedDRA GI Perforations SMQ (SMQ 20000107); note that this SMQ holds only narrow terms and has no broad terms.

A by-patient listing for all patients having a GI perforation event at any time will be provided and will include the information whether or not the event is considered as SAE, its relatedness to study drug, and whether or not it led to discontinuation from the study, led to permanent study drug discontinuation, or led to temporary study drug interruption.

If the number of events in the study warrants a table summary, the number and percent of patients with each PT will be provided, ordered by decreasing frequency in the baricitinib group (or combined baricitinib group, where appropriate). Additional summaries within this analysis will give an overview of events related to GI perforations. First, the total number of patients with GI perforations which are SAEs by ICH GCP or by protocol, by ICH GCP, and by protocol only, will be provided. Second, the number of patients with TEAEs associated with GI perforations that were rated as severe, and that were considered related to study drug in the opinion of the investigator, will be provided. Third, the number of patients with TEAEs associated with GI perforations that led to discontinuation from the study, that led to permanent study drug discontinuation, and that led to temporary study drug interruption will be provided.

6.14.4.11. Symptoms of Depression (QIDS-SR$_{16}$)

The Quick Inventory of Depressive Symptomatology Self-Rated (QIDS-SR$_{16}$) is a 16-item, self-report instrument intended to assess the existence and severity of symptoms of depression as listed in the American Psychiatric Association's Diagnostic and Statistical Manual of Mental Disorders, 4th Edition (DSM-IV; APA 1994). Patients are asked to consider each statement as it relates to the way they have felt for the past 7 days. There is a unique 4-point ordinal scale for each item with scores ranging from 0 to 3 reflecting increasing depressive symptoms as the item score increases. Additional information and the QIDS-SR$_{16}$ questions may be found on the University of Pittsburgh Epidemiology Data Center web site (http://www.ids-qids.org/index.html). A key reference for this instrument covers its psychometric properties for use in patients with chronic major depression (Rush et al. 2003).

The QIDS-SR$_{16}$ total score is derived as the sum of the scores across the 9 scale domains. For scale domains that contain more than 1 item, the domain score is the highest item rating given across the items within that domain):

- Sleep: the highest score on any one of the 4 sleep items (Items 1 to 4)
- Depressed mood (Item 5)
- Weight/appetite change: the highest score on any one of the 4 weight items (Items 6 to 9)
- Psychomotor changes: the highest score on either of the 2 psychomotor items (Items 15 and 16)
- Concentration (Item 10)
- Worthlessness/guilt (Item 11)

- Suicidal ideation (Item 12)
- Decreased interest (Item 13)
- Decreased energy (Item 14)

In the presence of missing data, the following rules will be employed to derive the total score. First, considering the 3 multi-item domains (sleep, weight/appetite change, psychomotor change), the domain score should be based on the maximum value across the appropriate items, and it should be missing only if each item is missing. Further, considering the 9 domain scores, the total score should be derived as missing if there are 3 or more domains that are missing; if 1 or 2 domain scores are missing, then the total score should be derived using a total score that is pro-rated to the full scale range (0 to 27) based on the available domain scores, retaining 1 decimal place in the total score derived in the presence of missing data.

The QIDS-SR$_{16}$ total scores will also be categorized in the following severity classes as shown in Table 6-8.

Table 6-8. QIDS-SR$_{16}$ Severity of Depressive Symptoms Categories based on the QIDS-SR$_{16}$ Total Score

QIDS-SR$_{16}$ Severity of Depressive Symptoms Category	QIDS-SR$_{16}$ Total Score
0=None	0-5
1=Mild	6-10
2=Moderate	11-15
3=Severe	16-20
4=Very Severe	21-27

Abbreviations: QIDS-SR$_{16}$ = Quick Inventory of Depressive Symptomatology Self-Rated-16.

Treatment differences in mean change in the QIDS-SR$_{16}$ total score may be analyzed by time point using an ANCOVA model with explanatory terms for treatment and the baseline value as a covariate. Treatment comparisons will be provided showing the LS means for each treatment, the treatment differences in LS means, the 95% CI for treatment differences, within-treatment p-values, and p-values for the treatment differences obtained from the model. Summary statistics for percent change from baseline will also be presented. Missing data will be imputed using mLOCF imputation.

Using the QIDS-SR$_{16}$ Severity of Depressive Symptoms Categories shown in Table 6-8, shift tables will show the number and percent of patients with total score in each category based on baseline to maximum category during the treatment period, with baseline depicted by the most extreme category during the baseline period, with further summarization of change from baseline in severity using categories of any improvement, no change, and any worsening, by treatment. Similarly, shift tables will be created for the QIDS-SR$_{16}$ suicidal ideation item (Item 12) responses. Treatment differences for the shift table summary categorizations will be assessed with the Likelihood Ratio Chi-Square test, using the exact version of this test whenever at least 1 cell has an expected cell count <5. Scheduled and unscheduled measurements will be included in these analyses. Furthermore, a summary of increases by each maximum severity category will be provided.

In addition, the number and percent of patients with total score in each category based on baseline to each scheduled visit, with baseline depicted by the last category during the baseline period, will be further summarized in terms of shift tables by time point and last visit and by treatment group. Following the shift table for the last visit, a shift table summary of the last visit will be included. Furthermore, a summary of increases by each category at last visit will be provided. Similarly, shift tables will be created for the QIDS-SR$_{16}$ suicidal ideation item (Item 12) responses by time point, including last visit.

Shift tables for the QIDS-SR$_{16}$ total score severity categories and for the suicidal ideation item descriptively comparing baseline to maximum and last post-treatment follow-up and follow-up baseline to maximum and last post-treatment follow-up by treatment group will be performed. A shift table summary for each analysis will also be included. Furthermore, a summary of increases by each maximum and last severity category will be provided.

First shift table summary:

- Decreased; maximum post-treatment follow-up category < baseline category
- Increased; maximum post-treatment follow-up category > baseline category
- Same; maximum post-treatment follow-up category = baseline category

Second shift table summary:

- Decreased; maximum post-treatment follow-up category < follow-up baseline category
- Increased; maximum post-treatment follow-up category > follow-up baseline category
- Same; maximum post-treatment follow-up category = follow-up baseline category

Third shift table summary:

- Decreased; last post-treatment follow-up category < baseline category
- Increased; last post-treatment follow-up category > baseline category
- Same; last post-treatment follow-up category = baseline category

Fourth shift table summary:

- Decreased; last post-treatment follow-up category < follow-up baseline category
- Increased; last post-treatment follow-up category > follow-up baseline category
- Same; last post-treatment follow-up category = follow-up baseline category

6.14.5. Study Drug Interruption

A summary of temporary interruptions of study drug will be provided, showing the number of patients who experienced at least 1 temporary interruption and the number of temporary interruptions per patient. Further, the duration of each temporary interruption (in days) and the cumulative duration of dose interruption (in days) using basic descriptive statistics (n, mean, SD, minimum, median, and maximum) will be displayed.

6.15. Subgroup Analyses

Subgroup analyses comparing baricitinib to placebo will be performed on the mITT population using ACR20, ACR50, DAS28-hsCRP <2.6, DAS28-hsCRP ≤3.2, SDAI ≤3.3, HAQ-DI Improvement ≥0.22, and change from baseline in HAQ-DI and DAS28-hsCRP at Week 12 (Visit 7) and Week 24 (Visit 11) and using mTSS at Week 24 (Visit 11).

The following subgroups (but not limited to only these) will be categorized into disease-related characteristics and demographic characteristics and will be evaluated:

- Gender (Male, Female)
- Age (<65 years, ≥65 years)
- Country
- Screening renal function status: impaired (eGFR <60 mL/min/1.73 m^2) or not impaired (eGFR ≥60 mL/min/1.73 m^2)
- Joint erosion status (1-2 joint erosions plus seropositivity, at least 3 joint erosions)
- Background therapy (MTX only, MTX + other cDMARDs)
- Baseline weight (≤median, > median)
- Baseline BMI category (≤median, >median)
- Baseline BMI category (<18.5 kg/m^2, ≥18.5 to <25 kg/m^2, ≥25 to <30 kg/m^2, ≥30 kg/m^2)
- Baseline disease severity (DAS28-hsCRP ≤5.1, DAS28-hsCRP >5.1)

The subgroup analyses for categorical outcomes will be performed using logistic regression. The model will include the categorical outcome based on NRI as the dependent variable and 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 logistic regression will be reported for the interaction test. When logistic regression sample size requirements (<5 responders in any category for any subgroup) are not met, the p-value will not be produced. Counts, percentages, the odds ratio, and the 95% CI of the odds ratio within a subgroup obtained from the same model with the subgroup and interaction terms removed will be presented for comparing treatment group differences. For significant findings, further analytical inspection will be made to assess the plausibility of any impact on drug effect.

An ANCOVA will be used to analyze the change from baseline based on mBOCF in HAQ-DI and DAS28-hsCRP at Week 12 and at Week 24, and based on linear extrapolation in mTSS at Week 24 with baseline, treatment, subgroup, and treatment-by-subgroup as explanatory variables. The treatment-by-subgroup interaction will be tested at the 0.1 significance level. The LS mean, LS mean difference, SE, and 95% CI within subgroups obtained from the same ANCOVA model with the interaction and subgroup terms removed will be presented.

Descriptive statistics will be provided for each treatment and stratum of a subgroup as outlined, regardless of sample size. In addition, the aforementioned formal statistical analysis will be performed when there are a minimum of 30 patients per treatment group per stratum of the subgroup. In some cases, select strata within the subgroups may need to be pooled in order to achieve 30 patients per treatment group per stratum for analysis. The descriptive and analysis statistics for the pooled strata will be displayed in addition to the descriptive statistics for the initial strata. The detailed pooling strategy is described below:

- For subgroups with 2 strata: gender, age, screening renal function status, background therapy, joint erosion status, baseline weight, baseline BMI category, baseline disease severity (DAS28-hsCRP) —If 1 of the strata fails to meet the required minimum number, then no formal statistical analysis will be provided.
- Country (3 strata)—Pool patients from Argentina and Brazil if needed. If after pooling, 1 of the strata fails to meet the required minimum number, then no formal statistical analysis will be provided.
- Baseline body mass index (four strata) —If one of the strata fails to meet the required minimum number, the stratum will be pooled with the next contiguous heavier BMI stratum, except for the heaviest BMI stratum which will be pooled with the next heaviest BMI stratum. The pooling will continue until all strata including pooled strata meet the required minimum number.

6.16. Exploratory Analysis

- Comparison between baricitinib and placebo in change from baseline to Weeks 16, 24, and 52 in mTSS: The actual scores and change from baseline scores at Week 16, 24, and 52 in mTSS will be summarized. An ANCOVA model with baseline score, country/region, baseline joint erosion status, and treatment group in the model will be used to test the treatment difference between baricitinib and placebo in change from baseline to Week 24 in mTSS scores. Type III sums of squares for the LS means will be used for the statistical comparison; 95% CI will also be reported. The LS mean difference, SE, CI, and p-value will be presented. The linear extrapolation method described in Section 6.3 will be used. The LOCF approach described in Section 6.3 will be used as a sensitivity approach to the linear extrapolation method to impute missing data.

  Supportive analyses using the cumulative responder curves, in which patients who have discontinued their randomized treatment are defined as progressors across the spectrum of choices of analysis thresholds will be provided.

- Descriptive statistics for the proportion of patients with mTSS change ≤0 by treatment group at Weeks 16, 24, and 52 will be provided by visit and treatment group. Logistic regression will be conducted to analyze the proportion of patients with mTSS change ≤0 at Weeks 16, 24, and 52 between baricitinib and placebo. The model for the logistic regression will consist of county/region, baseline joint erosion status, and treatment as fixed effects. Summaries by treatment group will be provided for the proportion of patients who improved or experienced no change (mTSS ≤0), or worsened (mTSS >0).

  In addition, the proportion of patients with change in mTSS ≤SDC and mTSS ≤0.5 at Weeks 16, 24, and 52 will be analyzed in a similar manner as described for mTSS change ≤0. The SDCs at Weeks 16, 24, and 52 are defined as the smallest detectable change which is computed from the variability in Week 0 to Week 16, Week 24, or Week 52 change in scores assigned by the blinded imaging assessors, respectively (Bruynesteyn 2005).

The linear extrapolation method described in Section 6.3 will be used to impute missing data. The categorical analyses described above will be applied to the mITT population, and a subset of mITT population that includes all randomized patients who received at least 1 dose of study drug, have non-missing baseline and at least 1 non-missing postbaseline mTSS measurement.

- Descriptive statistics for the actual score and change from baseline scores at Week 16, Week 24, and Week 52 in joint space narrowing and in bone erosion scores will be provided by visit and treatment group.
- Descriptive statistics for the proportion of patients achieving ACR20, ACR50, and ACR70 at each postbaseline visit through Visit 15 (Week 52) will be provided by visit and treatment group. A logistic regression with the same model specification as the primary analysis will be used for the treatment comparisons through Week 24 for ACR20, ACR50, and ACR70. The observed ("as-is") values of the ACR20, ACR50, and ACR70 will be summarized as a sensitivity approach to the primary method of NRI. Using the "as-is" approach, the ACR values are calculated without applying any imputation to the components.
- Descriptive statistics for the proportion of patients achieving EULAR Responder Index based on the 28-joint count at each postbaseline visit will be provided by visit and treatment group. EULAR nonresponders and responders (moderate + good responders) will be summarized. The NRI method will be used as the missing data imputation technique. A logistic regression analysis will be performed to compare the proportion of EULAR responders between baricitinib and placebo through Week 24. The model will consist of country/region, baseline joint erosion status, and treatment as fixed effects. The analysis will be repeated to compare the treatment groups in proportion of patients achieving good responses. Apart from the aforementioned categories, a summary table with frequencies and percentages will be presented for the patients categorized as good responder, moderate responder, and nonresponder. The observed values will be analyzed as a sensitivity approach to the primary method that utilizes NRI.
- A logistic regression analysis will be conducted to analyze the patients with specified categories up to Week 24 as below. The model for the logistic regression will consist of country/region, baseline joint erosion status, and treatment as fixed effects to compare baricitinib to placebo. Besides, descriptive statistics will be provided by visit and by treatment group for the following endpoints.
    - The proportion of patients achieving DAS28-ESR ≤3.2 and <2.6 and DAS28-hsCRP ≤3.2 and <2.6 at post-baseline visits.
    - The proportion of patients achieving HAQ-DI improvement from baseline ≥0.22 and ≥0.3
    - The proportion of patients achieving SDAI score ≤3.3 (ACR/EULAR remission index-based definition) and CDAI score ≤2.8 up to Week 52; the proportion of patients achieving CDAI score ≤10 and proportion of patients achieving SDAI score ≤11 up to Week 52.
    - The proportion of patients achieving ACR/EULAR remission according to the Boolean-based definition by treatment group.

- Change from baseline to postbaseline visits up to Week 24 in specific endpoints will be analyzed using an ANCOVA model with treatment, country/region, baseline joint erosion status, and baseline score in the model to test the treatment difference between baricitinib and placebo.

  Besides, descriptive statistics will be provided by visit and by treatment group for the following endpoints. The mLOCF approach described in Section 6.3.1 will be used to impute missing data.

  - The percent change from baseline and change from baseline in individual components of the ACR core set, including TJC68, SJC66, patient's assessment of arthritis pain, Patient's Global Assessment of Disease Activity, Physician's Global Assessment of Disease Activity, and hsCRP at postbaseline visits.
  - The actual score and the change in DAS28-hsCRP and in DAS28-ESR from baseline to post-baseline visits.
  - The actual score and the change in HAQ-DI from baseline to post-baseline visits.
  - The actual score and the change in CDAI and in SDAI from baseline to post-baseline.

- Descriptive statistics for the hybrid ACR (bounded) response at all postbaseline visits will be provided by visit and treatment group. An ANOVA model with country/region, baseline joint erosion status, and treatment group as fixed effects and the postbaseline hybrid ACR response as the dependent variable will be used to compare baricitinib to placebo through Week 24. The mLOCF approach described in Section 6.3.1 will be used to impute missing data

6.16.1.1. Analysis of Patient-Reported Outcomes Non-Diary Data

Observed patient-reported outcomes in duration of morning joint stiffness, Worst Tiredness NRS, and Worst Joint Pain NRS assessed using an electronic patient-reported outcomes. (ePRO) tablet by visit and ePRO change from baseline values at Weeks 12, 24, and 52 will be summarized in listings.

The effect of baricitinib compared to placebo from baseline up to Week 24 and to Week 52 in terms of descriptive summaries with regard to PROs will be evaluated by the following measures:

- **Duration of Morning Joint Stiffness**

  The duration of morning joint stiffness is a patient-administered item that allows for the patients to enter the length of time in minutes that their morning joint stiffness lasted on the day prior to that visit using an ePRO tablet. Durations recorded as longer than 12 hours (720 minutes) will be truncated to 720 minutes for analysis.

  Duration of morning joint stiffness will be measured at baseline and at postbaseline visits. Summary statistics of the actual score and change from baseline will be displayed by treatment and protocol-specified visit.

  The comparison between baricitinib and placebo in change from baseline through Week 24 in duration of morning joint stiffness will be analyzed using a nonparametric method involving the Hodges-Lehmann estimator. For the comparison between baricitinib and placebo, the median difference of change from baseline score and its 95% CI from the Hodges-Lehmann estimator will be presented. The p-value for the median difference of change from baseline will be calculated using the Wilcoxon rank-sum test. A 95% CI for the median change from baseline score within each treatment group will be calculated using distribution-free confidence limits. The mLOCF approach will be used to impute missing data. The observed values may be analyzed as a sensitivity approach to the primary method that utilizes mLOCF.

- **Tiredness Severity Numeric Rating Scale (Worst Tiredness NRS)**

  The Worst Tiredness NRS is a patient-administered, single-item, 11-point horizontal scale anchored at 0 and 10, with 0 representing "no tiredness" and 10 representing "as bad as you can imagine." Patients rate their tiredness by selecting the 1 number that describes their worst level of tiredness during the past 24 hours using an ePRO tablet.

  Worst Tiredness NRS will be measured at baseline and at postbaseline visits. Summary statistics of the actual score and change from baseline will be displayed by treatment and protocol-specified visit.

  The change from baseline in Worst Tiredness NRS to Week 24 will be analyzed using ANCOVA. The ANCOVA model will consider treatment, country/region, and baseline joint erosion status as fixed effects and baseline score as a covariate with change from baseline being the dependent variable. To test the treatment difference between baricitinib and placebo, the LS mean difference, 95% CI, and p-value will be presented. The mLOCF approach will be used to impute missing data. The observed values may be analyzed as a sensitivity approach to the primary method that utilizes mLOCF.

- **Severity of Joint Pain Numeric Rating Scale (Worst Joint Pain NRS)**

  The Worst Joint Pain NRS is a patient-administered, single-item, 11-point horizontal scale anchored at 0 and 10, with 0 representing "no pain" and 10 representing "pain as bad as you can imagine." Patients rate their pain by selecting the 1 number that describes their worst level of pain during the past 24 hours using an ePRO tablet.

  Worst Joint Pain NRS will be measured at baseline and at postbaseline visits. Summary statistics of the actual score and change from baseline will be displayed by treatment and protocol specified visit.

  The change from baseline in Worst Joint Pain NRS to Week 24 will be analyzed using ANCOVA. The ANCOVA model will consider treatment, country/region, and baseline joint erosion status as fixed effects and baseline score as a covariate with change from baseline being the dependent variable. To test the treatment difference between baricitinib and placebo, the LS mean difference, 95% CI, and p-value will be presented. The mLOCF approach will be used to impute missing data. The observed values may be analyzed as a sensitivity approach to the primary method that utilizes mLOCF.

- **FACIT-F**

  The FACIT-F scale (Cella and Webster 1997) is a 13-item, symptom-specific questionnaire that specifically assesses the self-reported severity of fatigue and its impact upon daily activities and functioning. The FACIT-F uses 0 ("not at all") to 4 ("very much") numeric rating scales to assess fatigue and its impact in the past 7 days. Scores range from 0 to 52 with higher scores indicating more fatigue.

FACIT-F scores will be measured at baseline and at postbaseline visits.

Summary statistics of the actual score and change from baseline will be displayed by treatment and protocol-specified visit.

The change from baseline in FACIT-F score to Week 24 will be analyzed using ANCOVA. The ANCOVA model will consider treatment, country/region, and baseline joint erosion status as fixed effects and baseline score as a covariate with change from baseline being the dependent variable. To test the treatment difference between baricitinib and placebo, the LS mean difference, 95% CI, and p-value will be presented. The mLOCF approach will be used to impute missing data.

The Minimum Clinically Important Difference (MCID) for the FACIT-F score is defined as ≥3.56 improvement (increase) from the baseline value of the Fatigue Subscale Score. The summary and analysis of the FACIT-F score will include the number and percent of patients achieving MCID at each postbaseline visit. The proportion of patients achieving the FACIT-F score MCID may also be analyzed using a logistic regression model with treatment, country/region, and baseline joint erosion status as fixed effects. The p-value and 95% CI of the odds ratio from the logistic regression model will be reported. The 95% CI of MCID rate difference will be calculated using the Newcombe-Wilson method without continuity correction. The NRI method described in Section 6.3.1 will be used to impute missing data.

- **WPAI-RA**

The WPAI-RA questionnaire was developed to measure the effect of general health and symptom severity on work productivity and regular activities in the last 7 days of the visit. It contains 6 items covering overall work productivity (health), overall work productivity (symptom), impairment of regular activities (health), and impairment of regular activities (symptom). Scores are calculated as impairment percentages (Reilly et al. 1993).

The WPAI-RA yields 4 types of scores:

- Absenteeism (work time missed)
- Presenteeism (impairment at work/reduced on-the-job effectiveness)
- Work productivity loss (overall work impairment/absenteeism plus presenteeism)
- Activity Impairment

Six questions will be answered:

- Q1 = currently employed
- Q2 = hours missed due to specified problem
- Q3 = hours missed due to other reasons
- Q4 = hours actually worked
- Q5 = degree that RA affected productivity while working
- Q6 = degree that RA affected regular activities

WPAI-RA outcomes are expressed as impairment percentages, with higher numbers indicating greater impairment and less productivity, ie, worse outcomes, as follows:

- Percentage work time missed due to RA “absenteeism”: Q2/(Q2+Q4)

- Percentage impairment while working due to RA "presenteeism": Q5/10
- Percentage overall work impairment due to RA "work productivity loss": Q2/(Q2+Q4)+[(1-Q2/(Q2+Q4))x(Q5/10)]
- Percentage activity impairment due to RA "activity impairment": Q6/10
- Multiply scores by 100 to express in percents.

WPAI-RA scores will be measured at baseline and at postbaseline visits.

The proportion of patients who were employed versus not employed at baseline will be summarized, along with by-visit summaries of changes in employment status (i.e. proportion of patients employed at baseline who are still employed at the postbaseline visit and the proportion of patients not employed at baseline who remained unemployed at the postbaseline visit). The analyses of absenteeism, presenteeism, and work productivity loss will be based on patients employed at both baseline and postbaseline visits. The analysis of activity impairment will be based on the mITT population. Summary statistics of the actual and change from baseline scores based on absenteeism, presenteeism, work productivity loss, and activity impairment will be displayed by treatment and protocol-specified visit.

The change from baseline in each of the 4 types of scores of WPAI-RA (absenteeism, presenteeism, work productivity loss, and activity impairment) to Week 24 may be analyzed using ANCOVA. The ANCOVA model will consider treatment, country/region, and baseline joint erosion status as independent fixed effects and the baseline score as a covariate with change from baseline being the dependent variable. To test the treatment difference between baricitinib and placebo, the LS mean difference, 95% CI, and p-value will be presented.

- **European Quality of Life–5 Dimensions–5 Level (EQ-5D-5L)**

  The EQ-5D-5L is a standardized measure of health status of the patient at the visit (same day) that provides a simple, generic measure of health for clinical and economic appraisal. The EQ-5D-5L consists of 2 components: a descriptive system of the respondent's health and a rating of his or her current health state using a 0- to 100-mm VAS. The descriptive system comprises the following 5 dimensions: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. Each dimension has 5 levels: no problems, slight problems, moderate problems, severe problems, and extreme problems. The respondent is asked to indicate his/her health state by ticking (or placing a cross) in the box associated with the most appropriate statement in each of the 5 dimensions. It should be noted that the numerals 1 to 5 have no arithmetic properties and should not be used as a cardinal score. The VAS records the respondent's self-rated health on a vertical VAS in which the endpoints are labeled "best imaginable health state" and "worst imaginable health state." This information can be used as a quantitative measure of health outcome. The EQ-5D-5L health states, defined by the EQ-5D-5L descriptive system, may be converted into a single summary index by applying a formula that essentially attaches a value (also called weights) to each of the levels in each dimension (Brooks 1996; EuroQol Group 2011 [WWW]; Herdman et al. 2011).

  The EQ-5D-5L questionnaire is completed by the patient in this study at baseline and post-baseline visits up to Week 52 (Visit 15). The EQ-5D-5L will be scored according to the developer's instructions (scoring guidelines). Three outcomes are described for each treatment group by visit according to the following:

- A health profile: summary statistics will be provided, including number of patients and proportions of categorical responses for the 5 dimensions (mobility, self-care, usual activity, pain/discomfort, and anxiety/depression)
- A health state index score: the UK algorithm (Szende et al. 2006) will be used to produce a patient-level index score between -0.59 and 1.0 (continuous variable); the same will be done using the US algorithm which produces a patient-level index score between -0.11 and 1.0.
- A self-perceived current health score: calculated from patient level EQ VAS responses (continuous variable).

The change in health state index scores and EQ VAS score from baseline through Week 24 (Visit 11) may be analyzed using an ANCOVA model with the following terms: treatment, country/region, baseline joint erosion status, and baseline score. Within each treatment arm, the LS mean, SE of the LS mean, and p-value will be reported. The LS mean difference between baricitinib and placebo, 95% CI for the mean difference, and p-value will be reported. ANCOVA will be conducted after imputing the postbaseline missing values using mLOCF.

All the aforementioned health outcome analyses will be performed on the mITT population. Descriptive statistics will be provided for all the parameters by visit and treatment group for baseline and all postbaseline visits.

- **Healthcare Resource Utilization**

The healthcare resource utilization data will be collected by site staff regarding the number of visits to medical care providers such as general practitioners, specialists, physical or occupational therapists, and other non-physical care providers for services outside of the clinical study; emergency room admissions; hospital admissions; and concomitant medications related to the treatment of RA. These data will be collected to support economic evaluations of treatment. A summary table including the following will be presented for the period of the past 2 months prior to study treatment as baseline and summarized at baseline and at post-baseline intervals (Week 0 to Week 12, Week 0 to Week 24, Week 24 to Week 52, and Week 0 to Week 52).

Healthcare resource utilization – Healthcare providers:

A summary of the number of consultations of medical care providers for care/treatment of RA will be provided for the period of the past 2 months prior to the study treatment. It will also be summarized at the post-baseline intervals described earlier. The number of consultations will be totaled by days and patients for the categories:

- Healthcare Provider Consultation
- Emergency Room Consultation

The "Healthcare Provider Consultation" comprises the following:

- General Practitioner (Such as family doctor or primary care physician)
- Specialist

- Rheumatologist
- Occupational Therapist
- Physical therapist
- Non-physician care providers

For the “Emergency Room Consultation” every necessary Emergency Room (ER) or equivalent facility consultation will be counted.

Healthcare resource utilization – Inpatient Hospitalization:

A summary table including the following will be presented from the period of the past 2 months prior to study treatment and at post-baseline intervals for the following details:

- Frequency of patient’s hospitalization and need for staying overnight because of their RA.
- Frequency of hospitalization type (general ward/intensive care ward).
- Descriptive statistics for the total number of days in hospital.

Healthcare resource utilization – Work Status:

A frequency table for the non-missing statements on work status for the period of the past 2 months prior to study treatment and at post-baseline intervals will be provided for the following details:

- Working for pay full-time (>36 hours per week)
- Working for pay part-time (≤36 hours per week)
- Unemployed, unrelated to study disease disability
- Unemployed due to study disease disability
- Other [Keeping house (full-time), Student (full-time), Student (part-time), Retired, Volunteer work]

6.17. Interim Analysis and Data Monitoring

No data monitoring committee is planned for the study.

No interim analysis is planned. A blinded data review will occur when all patients have completed the double-blind treatment period (Part A) for the primary endpoint. Pooled primary endpoint, secondary endpoints, and safety measurements will be provided for study team review for the purpose of data cleaning and understanding. No decision related to sample size adjustment, efficacy, and inferiority will be made.

7. Unblinding Plan

There will be 2 planned database locks for this study: the first database lock will occur after all patients complete Part A (Week 24) and a final database lock at the end of the study. Only 1 unblinding will occur after the final database lock.

At the first database lock, by-visit data through Week 24 (Visit 11) will be locked but ongoing log-type entry forms for ongoing patients such as AEs and concomitant medications will remain unlocked. Blinded analyses will be conducted on the data that are collected through Week 24. X-ray results will only be analyzed after final database lock.

A final database lock is planned at the end of the study when all patients complete the study. The final analysis will include all data that are collected during the study. Sponsor, all investigators, study site personnel, and patients will remain blinded to treatment assignments until the final database lock.

8. References

[APA] American Psychiatric Association. Diagnostic and Statistical manual of Mental Disorders. 4th ed. DSM-IV. Washington, DC: American Psychiatric Association; 1994.

Brooks R. EuroQol: The current state of play. *Health Policy.* 1996;37(1):53–72.

Bruce B and Fries J. The Standaford Health Assessment Questionnaire (HAQ): A review of its history, issues, progress and documentation. *J Rheumatol.* 2003;30(1):167-178.

Bruynesteyn K, Boers M, Kostense P, van der Linden S, van der Heijde D. Decidingon progression of joint damage in paired films of individual patients: smallest detectable difference or change. *Ann Rheum Dis.* 2005;64(2):179-182.

Cancer Therapy Evaluation Program, Common Terminology Criteria for Adverse Events, Version 3.0, DCTD, NCI, NIH, DHHS, March 31, 2003.

Cohen SB, Emery P, Greenwald MW, Dougados M, Furie RA, Genovese MC, Keystone EC, Loveless JE, Burmester GR, Cravets MW, Hessey EW, Shaw T, Totoritis MC. Rituximab for rheumatoid arthritis refractory to anti-tumor necrosis factor therapy. Results of a multicenter, randomized, double-blind, placebo-controlled phase III trial evaluating primary efficacy and safety at twenty-four weeks. *Arthritis Rheum.* 2006;54(9):2793-2806.

Common Terminology Criteria for Adverse Events, Version 4.03, DHHS NIH NCI, NIH Publication No. 09-5410, June 14, 2010.

D'Agostino RB, Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM, Kannel WB. A General Cardiovascular Risk Profile for Use in Primary Care: The Framingham Heart Study. *Circulation.* 2008;117(6):743-753.

The EuroQol Group. EQ-5D-5L User Guide. Version 1.0. April 2011. Available at: http://www.euroqol.org/fileadmin/user_upload/Documenten/PDF/Folders_Flyers/UserGuide_EQ-5D-5L.pdf.

Fries JF. The Assessment of disability: from first to future principles. *Br J Rheumatol.* 1983;22(3 Suppl):48-58.

Fries JF. The hierarchy of quality-of-life assessment, the Health Assessment Questionnaire (HAQ), and issues mandating development of a toxicity index. *Control Clin Trials.* 1991;12:106S-117S.

Fries JF, Spitz P, Kraines RG, Holman HR. Measurement of patient outcome in arthritis. *Arthritis Rheum*. 1980;23(2):137-145.

Fries JF, Spitz PW, Young DY. The dimensions of health outcomes: the health assessment questionnaire, disability and pain scales. *J Rheumatol.* 1982;9(5):789-793.

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

Keystone EC, Kavanaugh AF, Sharp JT, Tannenbaum H, Hua Y, Teoh LS, Fischkoff SA, Chartash EK. Radiographic, clinical, and functional outcomes of treatment with adalimumab (a human anti–tumor necrosis factor monoclonal antibody) in patients with active rheumatoid arthritis receiving concomitant methotrexate therapy: a randomized, placebo-controlled, 52-week trial. *Arthritis Rheum.* 2004;50(5):1400-1411.

Keystone E, van der Heijde D, Mason D Jr, Landewé R, van Vollenhoven R, Combe B, Emery P, Strand V, Mease P, Desai C, Pavelka K. Certolizumab pegol plus methotrexate is significantly more effective than placebo plus methotrexate in active rheumatoid arthritis: findings of a fifty-twoweek, phase III, multicenter, randomized, double-blind, placebo-controlled, parallel-group study. *Arthritis Rheum.* 2008;58(11):3319-3329.

Keystone E, Emery P, Peterfy CG, Tak PP, Cohen S, Genovese MC, Dougados M, Burmester GR, Greenwald M, Kvien TK, Williams S, Hagerty D, Cravets MW, Shaw T. Rituximab inhibits structural joint damage in patients with rheumatoid arthritis with an inadequate response to tumour necrosis factor inhibitor therapies. *Ann Rheum Dis.* 2009;68(2):216-221.

National Kidney Foundation. K/DOQI Clinical Practice Guidelines for Chronic KidneyDisease: Evaluation, Classification and Stratification. *Am J Kidney Dis.* 20023; 39(2)(suppl 1):S1-S266.

Newcombe RG. Interval estimation for the difference between independent proportions: comparison of eleven methods. *Stat Med.* 1998;17:873-890. Erratum in Stat Med. 1999;18:1293.

Ramey DR, Fries JF, Singh G. The Health Assessment Questionnaire 1995-status and review. In: Quality of Life and pharmacoeconomics in clinical trials. Second edition. Edited by B Spilker. Lippincott-Raven Publishers, Philadelphia, 1996

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

Ridker PM, Buring JE, Rifari N, Cook NF. Development and Validation of Improved Algorithms for the Assessment of Global Cardiovascular Risk in Women: The Reynolds Risk Score. *JAMA.* 2007;(297):6.

Ridker PM, Paynter NP, Rifai N, Gaziano JM, Cook NR. C-Reactive Protein and Parental History Improve Global Cardiovascular Risk Prediction: The Reynolds Risk Score for Men. *Circulation.* 2008;118:2243-2251.

Rush AJ, Trivedi MH, Ibrahim HM, Carmody TJ, Arnow B, Klein DN, Markowitz JC,Ninan PT, Kornstein S, Manber R, Thase ME, Kocsis JH, Keller MB. The 16-item Quick Inventory of Depressive Symptomatology (QIDS) Clinician Rating (QIDS-C) and Self-Report (QIDS-SR): A psychometric evaluation in patients with chronic major depression. *Biological Psychiatry.* 2003;54:573-583.

Sampson, HA, Muñoz-Furlong, A, Campbell RL, Adkinson NF Jr, Bock SA, Branum A, Brown SG, Camargo CA Jr, Cydulka R, Galli SJ, Gidudu J, Gruchalla RS, Harlor AD Jr, Hepner DL, Lewis LM, Liberman PL, Metcalfe DD, O'Connor R, Muraro A, Rudman A, Schmitt C, Scherrer D, Simons FE, Thomas S, Wood JP, Decker WW. Second symposium on the definition and management of anaphylaxis: summary report - second National Institute of Allergy and Infectious Disease/Food Allergy and Anaphylaxis Network symposium. *J Allergy Clin Immunol.* 2006;117(2):391-397.

Smolen J, Landewé RB, Mease P, Brzezicki J, Mason D, Luijtens K, van Vollenhoven RF, Kavanaugh A, Schiff M, Burmester GR, Strand V, Vencovskỳ J, van der Heijde D. Efficacy and safety of certolizumab pegol plus methotrexate in active rheumatoid arthritis: the RAPID 2 study. A randomised controlled trial. *Ann Rheum Dis.* 2009;68(6):797-804.

Smolen JS, Beaulieu A, Rubbert-Roth A, Ramos-Remus C, Rovensky J, Alecock E, Woodworth T, Alten R, OPTION investigators. Effect of interleukin-6 receptor inhibition with tocilizumab in patients with rheumatoid arthritis (OPTION study): a double-blind, placebocontrolled, randomised trial. *Lancet.* 2008;371(9617):987–997.

Szende A, Oppe M, Devlin N. EQ-5D value sets: Inventory, comparative review and user guide. EuroQol Group Monographs. Volume 2. Springer, 2006.

Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III) Final Report, National Cholesterol Education Program, National Heart, Lung, and Blood Institute, National Institutes of Health, NIH Publication No. 02-5215, September 2002.

University of Pittsburgh IDS/QIDS page. IDS/QIDS web site. Available at: http://www.ids-qids.org. Accessed June 13, 2012.

van der Heijde D. How to read radiographs according to the Sharp/van der Heijde method. *J Rheumatol.* 2000; 27(1):261-263

9. Appendices

Appendix 1. Algorithm for Determining ACR Response

Details presented in this appendix will use “x” as a generic symbol, and the appropriate number (either 20, 50, or 70) is to be filled in when implementing in dataset programming code.

ACRχ response is defined as ≥χ% improvement from baseline in tender joint count (68 joint counts), and ≥χ% improvement in swollen joint count (66 joint counts), and ≥χ% improvement in at least three of the following five items:

- Patient’s assessment of pain;
- Patient’s Global Assessment of Disease Activity;
- Physician’s Global Assessment of Disease Activity;
- HAQ-DI;
- hsCRP.

The following abbreviations will be used throughout this appendix to refer to the items needed in the algorithm definitions:

Parameter	Abbreviation for the Parameter
% improvement in tender joint count	TJC68
% improvement in swollen joint count	SJC66
% improvement in patient’s assessment pain	PATPAIN
% improvement in Patient’ Global Assessment of Disease Activity	PATGA
% improvement in Physician’s Global Assessment of Disease Activity	PHYGA
% improvement in HAQ-DI	HAQ
% improvement in hsCRP	hsCRP

For all seven parameters mentioned above, % improvement at a visit is calculated as:
(baseline value – value at visit) x 100 / baseline value.

<u>To calculate the observed ACRx response at a visit:</u>

- Step 1: If the patient discontinued from the study prior to reaching the visit, then STOP – assign ACRx response as blank (ie, missing). Otherwise, calculate the % improvement at the visit for all seven parameters as described above.
- Step 2:
  - o If TJC68 AND SJC66 are BOTH ≥χ%, then proceed to step3.
  - o If both are non-missing but one or both is <χ%, then STOP – assign the patient as a non-responder for ACRχ.
  - o If either or both are missing, proceed as follows:
    a. If both are missing, then STOP – assign ACRx response as blank (ie, missing).
    b. If one of TJC68 or SJC66 is missing and the non-missing value is <χ%, then STOP – assign the patient as a non-responder for ACRχ.

    c. If one of TJC68 or SJC66 is missing and the non-missing value is ≥χ%, then STOP – assign ACRx response as blank (ie, missing).

- Step 3: Consider the following five variables: PATPAIN, PTGA, PHYGA, HAQ, and hsCRP.
  - o If three or more items are missing, then STOP – assign ACRx response as blank (ie, missing).
  - o If three or more items are non-missing, then proceed <u>with the following order</u>:
    a. If at least three items are ≥χ%, then STOP – assign the patient as a responder for ACRχ.
    b. If at least three items are <χ%, then STOP – assign the patient as a nonresponder for ACRχ.
    c. If less than three items are ≥χ%, then STOP – assign ACRχ response as blank.

<u>To calculate the ACRx response at post baseline visits up to Week 52 using NRI:</u>

- Step 1: Calculate the % improvement at the visit for all seven parameters as described above.
- Step 2: If all seven parameters are missing at the visit, or if the patient discontinued from the study prior to reaching the visit, then STOP – assign the patient as a nonresponder for ACRx.
- Step 3: If at least one of the seven parameters is non-missing at the visit and the patient is still enrolled in the study, then use LOCF to fill in any missing values.
- Step 4: If TJC68 AND SJC66 are BOTH ≥χ %, then proceed to step 5. If both are nonmissing but one or both is <χ %, then STOP – assign the patient as a nonresponder for ACR χ. If either or both are missing, then STOP – assign the patient as a non-responder for ACR χ.
- Step 5: Consider the following five variables: PATPAIN, PATGA, PHYGA, HAQ, hsCRP.
  - o If three or more of the five items are non-missing, then:
    - ▪ If three or more of the five items are ≥ χ%, then STOP – assign the patient as a responder for ACR*x*.
    - ▪ If less than three of those five items are ≥ χ%, then STOP – assign the patient as a non-responder for ACR*x*.
  - o If three or more of the five items are missing, then STOP – assign the patient as a non-responder for ACR*x*.

Appendix 2. Details of Efficacy and Health Outcomes Measures

van der Heijde Modified Total Sharp Score

X-rays of the hands/wrists and feet will be scored (according to the Bioclinica predefined imaging charter) for structural progression as measured using the mTSS (van der Heijde, 2000). This methodology quantifies the extent of bone erosions for 44 joints and joint space narrowing for 42 joints with higher scores representing greater damage.

Structural progression will be measured using the mTSS. During screening, all patients meeting entry criteria will have a single posteroanterior radiographic assessment of the left and right hand/wrist and a single dorsoplantar radiographic image taken of the left and right foot. Alternatively, images taken within 4 weeks prior to baseline may be used. Images taken at screening (or within 4 weeks of screening) will be reviewed centrally by qualified readers for evidence of erosive bony change(s). These initial radiographs will serve as the baseline radiographs for comparison throughout the study.

Additional radiographs (ie, a single posteroanterior view of each hand and a single dorsoplantar view of each foot) will be obtained at Week 16, Week 24 (the primary evaluation), and Week 52 (for the purposes of demonstrating durability of structural preservation), as shown in the study schedule (See Protocol Appendix). For patients who discontinue prematurely from the study, x-rays will be performed at the early termination visit only if the previous x-rays were taken more than 12 weeks earlier.

Bone Joint Erosion Score

The joint erosion score is a summary of erosion severity in 32 joints of the hands and 12 joints of the feet. The maximum erosion score for a hand joint is 5 and for a foot joint is 10. Thus, the maximal erosion score is 280 for a time point (160 for both hands/wrists and 120 for both feet). Each joint is scored according to the surface area involved from 0 to 5 for hand joints and 0 to 10 for the foot joints. The highest score (5 for the hand and 10 for the foot) indicates extensive loss of bone from more than one half of the articulating bone. A score of 0 in either the hand or foot joints indicates no erosion.

Erosion for each hand joint (16 joints per hand) is graded on the following scale.

0 – Normal (no erosion)

1 – Discrete small erosion

2 – Two discrete erosions or one large erosion not passing the joint middle line

3 – One large erosion passing the joint middle line of the bone or combination of the above

4 – Combinations of discrete and/or large erosions adding up to 4

5 – Combinations of discrete and/or large erosions adding up to 5 or more

NA - Not assessable due to poor radiographic depiction

S - Surgically modified (e.g. joint replacement, amputation, or other surgery)

Discrete erosions will be graded 1 if small and 2 or 3 if larger, with a score of 3 if the erosion is large and extends across the imaginary middle of the bone. Discrete erosions of each grade will then be summed over both sides of the joint to a maximum of 5 to give a total score for each location in the hands. When erosions in the carpal bones of the wrist are confluent rather than discrete and focal, the percent surface eroded will be scored from 0 to 5 in approximately 20% intervals. A score of 'NA' will be assigned in the case that the location is not evaluable due to advanced subluxation or poor radiographic depiction. A score of 'S' will be assigned in the case of joint replacement, amputation, or other surgery.

Erosion for each side of a foot joint (6 joints per foot) is graded on the following scale.

0 – Normal (no erosion)

1 – Discrete small erosion

2 – Two discrete erosions or one large erosion not passing the joint middle line

3 – One large erosion passing the joint middle line

4-10 – Combination of the above conditions.

NA - Not assessable due to poor radiographic depiction

S - Surgically modified (e.g. joint replacement, amputation, or other surgery)

Since both sides of each foot joint are graded, the maximum score for a foot joint is 10 (a score of 5 on each side).

Joint Space Narrowing (JSN) Score

The JSN score summarizes the severity of JSN in 30 joints of both hands and 12 joints of both feet. Assessment of JSN for each hand (15 joints per hand) and foot (6 joints per foot), including subluxation, is scored from 0 to 4, with 0 indicating no (normal) JSN and 4 indicating complete loss of joint space, bony ankylosis, or luxation. Thus, the maximum JSN score is 168 (120 for both hands/wrists and 48 for both feet) at a time point. JSN for each hand and foot joint is graded on the following scale:

0 – Normal or no narrowing

1 – Asymmetrical and/or minimal narrowing- up to 25%

2 – Definite narrowing with loss of up to 50% of the normal space

3 – Definite narrowing with loss of 50-99% of the normal space or subluxation

4 – Absence of a joint space, presumptive evidence of ankylosis, or complete luxation

NA - Not assessable due to poor radiographic depiction

S - Surgically modified (e.g. joint replacement, amputation, or other surgery)

Total Modified Sharp Score (mTSS)

The total mTSS score at a time point is the sum of the erosion (maximum of 280) and JSN (maximum of 168) scores, for a maximum score of 448.

Handling of Missing Joint Data

Refer to a separate document for a detailed description of imputation of missing joint data.

Adjudication Process

The read of x-ray images will be performed by 2 independent readers. An adjudication of results will be in place to show any discrepancy between the 2 independent readers. The adjudication process will involve re-reading by a third independent reader. Cases that require adjudication will be identified once the 2 primary independent reviews are completed. To ensure a consistent read by the 2 independent readers, an inter-/intra-reader variability assessment will be evaluated. Refer to the Bioclinica imaging charter and a separate document describing missing data imputation for more details.

ACR Core Set

The individual components that make up the ACR Core Set of measures for RA are:

Tender Joint Count (TJC)

For ACR measures, the number of tender and painful joints will be determined by examination of 68 joints (34 joints on each side of the patient’s body). These 68 joints will be assessed and classified as tender, not tender, or not evaluable. The number of tender joints ranges from 0 to 68. The algorithm to calculate TJC is in the Joint Assessment section of this Appendix.

Swollen Joint Count (SJC)

For ACR measures, the number of swollen joints will be determined by examination of 66 joints (33 joints on each side of the patient’s body). These 66 joints will be assessed and classified as swollen, not swollen, or not evaluable. The number of swollen joints ranges from 0 to 66. Refer to the Joint Assessment section of this Appendix for the algorithm to calculate SJC.

Patient’s Assessment of Pain (VAS)

The patient will be asked to assess his or her current level of pain by marking a vertical tick on a 100-mm horizontal VAS with the left end marked as “no pain” and the right end marked as “worst possible pain”. Results will be expressed in millimeters measured between the left end of the scale and the crossing point of the vertical line of the tick. This procedure is applicable for all VAS scales used in the trial.

Patient’s Global Assessment of Disease Activity (VAS)

The patients will be asked to give an overall assessment of how their rheumatoid arthritis is affecting them at present by marking a vertical tick on a VAS from “very well” to “very poor”.

Physician’s Global Assessment of Disease Activity (VAS)

The physician’s assessment of a patient’s disease activity assessed at the visit will be recorded using the VAS from “none” to “extremely active arthritis”.

Patient’s Assessment of Physical Function (HAQ-DI)

The patient will be asked to complete the HAQ-DI to measure the impact of arthritis on their ability to function in daily life. The HAQ-DI functional disability index score ranges from 0 to 3, with lower scores indicating better physical functioning. See Section HAQ-DI of this Appendix for the algorithm to calculate the HAQ-DI functional disability index score.

High sensitivity C-Reactive Protein (hsCRP)

hsCRP will be the ACR Core Set measure of acute phase reactant. It will be measured at the central laboratory.

Joint Assessment

Each of 28 or 68 joints will be evaluated for tenderness and 28 or 66 joints will be evaluated for swelling (hips are excluded for swelling) at the specified visits as shown in the schedule of events of the protocol. The 68/66 joint count will be performed at screening, Week 0 (baseline), Week 1 to Week 24 in Part A, Week 24 to Week 52 in Part B, end of treatment, and at the follow-up visit. For scores using 28 joints, the following subset will be assessed (both right and left side):

- Shoulder,
- Elbow,
- Wrist,
- Metacarpophalangeal (MCP) I, II, III, IV, and V,
- Thumb interphalangeal,
- Proximal interphalangeal (PIP) II, III, IV, and V,
- KJannee.

Joints will be assessed for tenderness by pressure and joint manipulation on physical examination. Any positive response on pressure, movement, or both will then be translated into a single tender-versus non-tender dichotomy. Joints will be classified as either swollen or not swollen. Swelling is defined as palpable fluctuating synovitis of the joint. Swelling secondary to osteoarthrosis will be assessed as not swollen, unless there is unmistakable fluctuation.

The following joint count imputation rules will be applied.

(1). Joints that had any of the following procedures or disease conditions that occur at the screening (Visit 1) up to and including baseline (Visit 2) will be imputed as follows:

- **Arthroplasty, fusion, synovectomy, ankylosis, amputation, injury, fracture, and infection:** these will be considered "non-evaluable" from baseline up to end of the study. Joints that include "non-evaluable" joints will be prorated from baseline up to end of the study. For example, the swollen joint count score for 6 swollen plus 2 "non-evaluable" joints for DAS28 is calculated as [6/(28-2)]*28 = 6.46.
- **Injection:** joints that have received intra-articular and bursal injections will be collected in the data but set to painful/tender or swollen after the injection for 6 months (up to Week 24 [Visit 11]).

(2). Joints that had any of the following procedures or disease conditions that occur postbaseline (anytime during treatment with the study drugs) will be imputed as follows:

- **Amputation:** amputation that occurs postbaseline will be considered "non-evaluable" from the visit after amputation up to end of the study. Joints that include "non-evaluable" joints will be prorated from the visit after amputation up to end of the study. For example, the swollen joint count score for 6 swollen plus 2 "non-evaluable" joints for DAS28 is calculated as [6/(28-2)]*28 = 6.46.

- **Arthroplasty, fusion, synovectomy, and ankylosis:** if any of these procedures occur postbaseline, joints will be set to painful/tender or swollen from the visit after the procedure up to end of the study.
- **Infection, injury, and fracture:** if any of these conditions occur postbaseline, joints will be set to painful/tender or swollen from the visit after the condition until the condition is resolved.
- **Injection:** joints that have received intra-articular and bursal injections postbaseline will be collected in the data but set to painful/tender or swollen after the injection for 6 months or end of study (whichever comes first).

The number of tender and swollen joints will be calculated by summing all joints. For patients who have an incomplete set of joints evaluated, the joint count will be adjusted to a 28- or 68-joint count for tenderness and a 28- or 66-joint count for swelling by dividing the number of affected joints by the number of evaluated joints and multiplying by 28 or 68 for tenderness and 28 or 66 for swelling.

Disease Activity Score-Erythrocyte Sedimentation Rate (DAS28-ESR) and Disease Activity Score-High Sensitivity C-Reactive Protein (DAS28-hsCRP)

The DAS28 is a measure of disease activity in 28 joints that consists of a composite numeric score of the following variables: TJC, SJC, hsCRP or ESR, and Patient's Global Assessment of Disease Activity (Vander Cruyssen et al. 2005). The 28 joints to be examined and assessed as tender or not tender for TJC and as swollen or not swollen for SJC include 14 joints on each side of the patient's body: the 2 shoulders, the 2 elbows, the 2 wrists, the 10 metacarpophalangeal joints, the 2 interphalangeal joints of the thumb, the 8 proximal interphalangeal joints, and the 2 knees (Smolen et al. 1995). The following equation will be used to calculate the DAS28- hsCRP (Vander Cruyssen et al. 2005):

$$\text{DAS28-hsCRP} = 0.56(\text{TJC28})^{1/2} + 0.28(\text{SJC28})^{1/2} + 0.36(\ln(\text{hsCRP} + 1)) + 0.014(\text{VAS}) + 0.96$$

The hsCRP value in mg/L must be used. In order to calculate DAS28-hsCRP data must be present for all 4 components: TJC28, SJC28, patient's global assessment (VAS), and hsCRP. Component-level imputation will be performed for the calculation of DAS28-hsCRP for the imputed visits summaries. If at least 1 of the 4 components is nonmissing at the visit and the patient is still enrolled in the study, then use LOCF to fill in any missing values.

Based on the DAS-hsCRP, the following will be calculated:

- DAS28-hsCRP ≤3.2
- DAS28-hsCRP <2.6

Similarly,

$\text{DAS28-ESR} = 0.56(\text{TJC28})^{1/2} + 0.28(\text{SJC28})^{1/2} + 0.70 \ln(\text{ESR}) + 0.014(\text{VAS})$ [If ESR = 0 in analysis data, set ln(ESR) = 0]

The ESR (Erythrocyte sedimentation rate) in mm/first hour must be used. In order to calculate DAS28-ESR data must be present for all 4 components: TJC28, SJC28, patient's global assessment (VAS), and ESR. Component-level imputation will be performed for the calculation of DAS28-ESR for the imputed visits summaries. If at least 1 of the 4 components is nonmissing at the visit and the patient is still enrolled in the study, then use LOCF to fill in any missing values.

Based on the DAS-ESR, the following will be calculated:

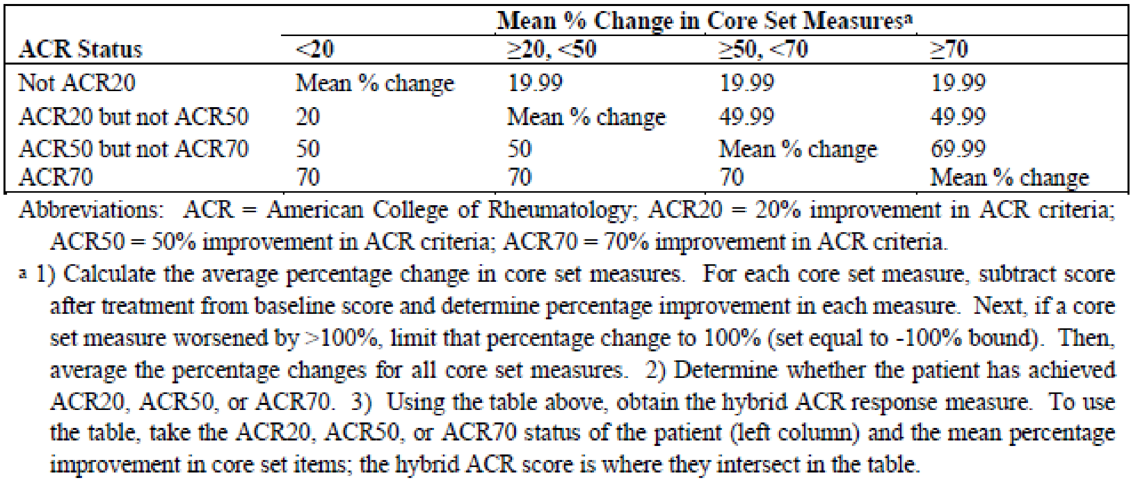
- DAS28-ESR ≤3.2
- DAS28-ESR <2.6

A NRI rule for the DAS28 binary endpoints will be applied for each visit as follows: if the DAS28 is missing at the visit or if the patient discontinued treatment or the study prior to the visit, then assign the patient as DAS28 non-responder.

Hybrid American College of Rheumatology Response Measure

The hybrid ACR (bounded) response measure will be obtained as described by the ACR Committee to Reevaluate Improvement Criteria (2007);

Table APP.1. Scoring Method for the Hybrid American College of Rheumatology Response Measure



	Mean % Change in Core Set Measures[a]			
ACR Status	<20	≥20, <50	≥50, <70	≥70
Not ACR20	Mean % change	19.99	19.99	19.99
ACR20 but not ACR50	20	Mean % change	49.99	49.99
ACR50 but not ACR70	50	50	Mean % change	69.99
ACR70	70	70	70	Mean % change

Abbreviations: ACR = American College of Rheumatology; ACR20 = 20% improvement in ACR criteria; ACR50 = 50% improvement in ACR criteria; ACR70 = 70% improvement in ACR criteria.

[a] 1) Calculate the average percentage change in core set measures. For each core set measure, subtract score after treatment from baseline score and determine percentage improvement in each measure. Next, if a core set measure worsened by >100%, limit that percentage change to 100% (set equal to -100% bound). Then, average the percentage changes for all core set measures. 2) Determine whether the patient has achieved ACR20, ACR50, or ACR70. 3) Using the table above, obtain the hybrid ACR response measure. To use the table, take the ACR20, ACR50, or ACR70 status of the patient (left column) and the mean percentage improvement in core set items; the hybrid ACR score is where they intersect in the table.

Simplified Disease Activity Index (SDAI)

The SDAI is a tool for measurement of disease activity in RA that integrates measures of physical examination, acute phase response, patient self-assessment, and evaluator assessment. The SDAI is calculated by adding together scores from the following assessments:

- number of swollen joints (0 to 28),
- number of tender joints (0 to 28),
- hsCRP in mg/dL (0.1 to 10.0),
- Patient’s Global Assessment of Disease Activity on VAS (0 to 10.0, measured in cm), and
- evaluator or Physician’s Global Assessment of Disease Activity on VAS (0 to 10.0, measured in cm) (Aletaha and Smolen 2005)

Disease remission according to ACR/EULAR index-based definition of remission is defined as an SDAI score of ≤3.3 (Felson et al. 2011).

Low disease activity according to SDAI is defined as a SDAI score ≤11 (Aletaha and Smolen 2005).

Clinical Disease Activity Index (CDAI)

The CDAI is similar to the SDAI, but it allows for immediate scoring because it does not use a laboratory result. The CDAI is calculated by adding together scores from the following assessments:

- number of swollen joints (0 to 28),
- number of tender joints (0 to 28),
- Patient Global Assessment of Disease Activity on VAS (0 to 10.0, measured in cm), and
- evaluator or Physician's Global Assessment of Disease Activity on VAS (0 to 10.0, measured in cm) (Aletaha and Smolen 2005)

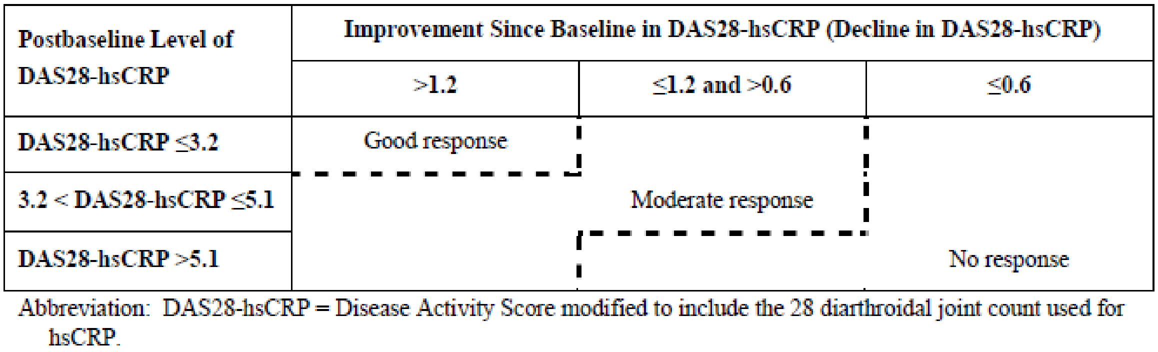
Remission according to CDAI is defined as a CDAI score of ≤2.8 (Felson et al. 2011).

Low disease activity according to CDAI is defined as a CDAI score ≤10 (Aletaha and Smolen 2005).

European League Against Rheumatism (EULAR) Responder Index

Assessments of patients with RA by the EULAR Responder Index based on the 28-joint count will be used to categorize patients as nonresponders, moderate responders, good responders, or responders (moderate + good responders) according to van Gestel et al. 1998.

Table APP.2. Categorization of Patients as Nonresponders, Moderate Responders, or Good Responders



Postbaseline Level of DAS28-hsCRP	Improvement Since Baseline in DAS28-hsCRP (Decline in DAS28-hsCRP)		
	>1.2	≤1.2 and >0.6	≤0.6
DAS28-hsCRP ≤3.2	Good response		
3.2 < DAS28-hsCRP ≤5.1		Moderate response	
DAS28-hsCRP >5.1			No response

Abbreviation: DAS28-hsCRP = Disease Activity Score modified to include the 28 diarthroidal joint count used for hsCRP.

There are 3 categories, no improvement (or no response), moderate improvement (or moderate response), and good improvement (highest kind of improvement or good response). The category of no improvement also includes a worsening of RA. Hence, relative to baseline, a patient can only have a "0" change in categories, a change of "1" category or a change of "2" categories. EULAR response based on DAS28-ESR may also be evaluated.

ACR/EULAR Rheumatoid Arthritis Remission

Two new ACR/EULAR definitions of RA remission will be evaluated, a “boolean-based definition” and an “index-based definition” (Felson et al. 2011).

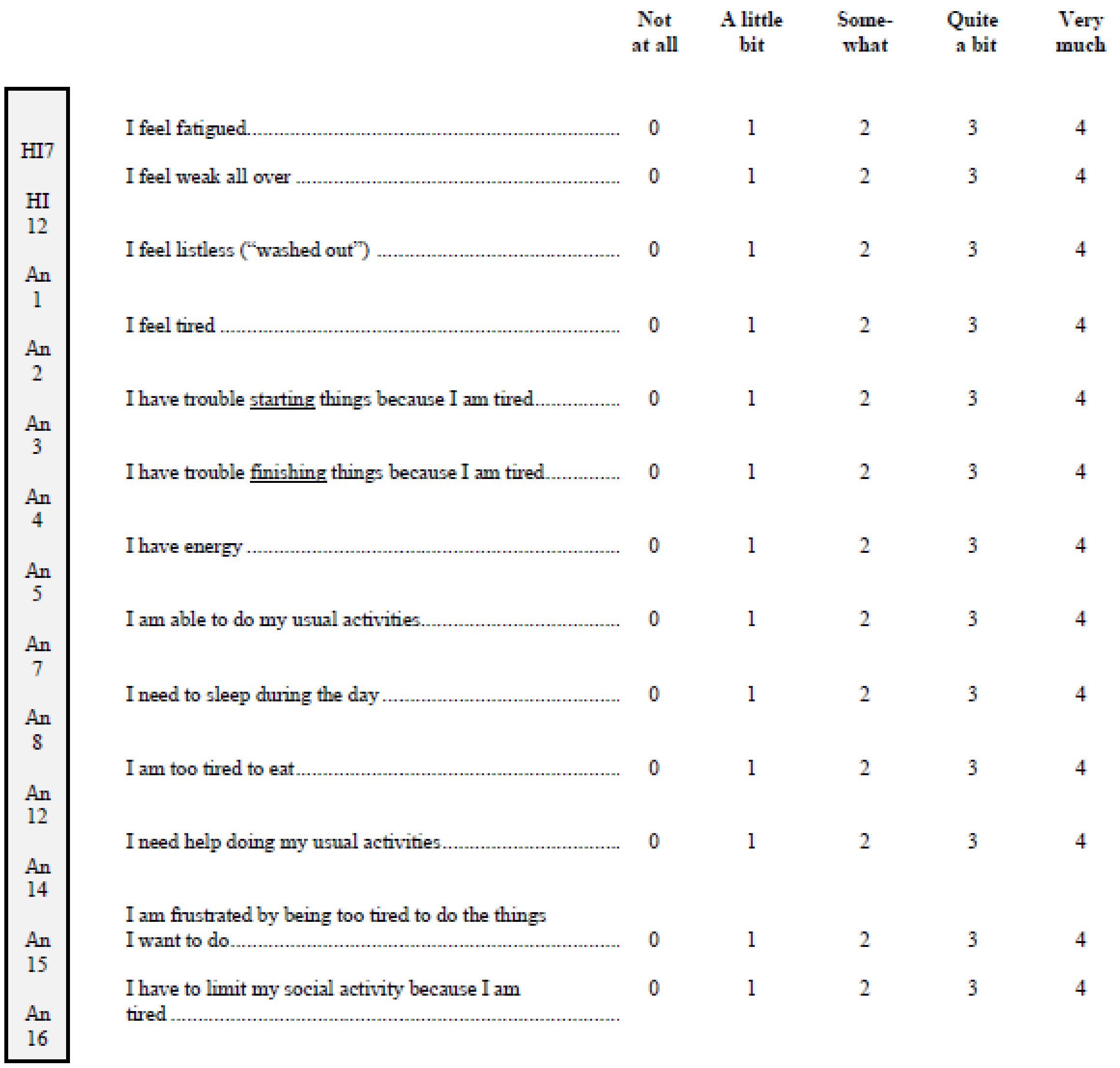
- Boolean-based Definition of Remission: All 4 criteria below must be met (at the same visit):
  - TJC28 ≤1
  - SJC28 ≤1
  - hsCRP ≤1 mg/dL (10 mg/L)
  - Patient’s Global Assessment of Disease Activity (on a 0-100 mm scale) / 10 ≤1
- Index-based Definition of Remission:
  - Simplified Disease Activity Index (SDAI) ≤3.3

Patient’s Global Assessment of Disease Activity / 10 in order to convert from a 0-100 mm scale to a 0-10 cm scale.

The hsCRP value converted to mg/dL will be used.

Functional Assessment of Clinical Illness Therapy Fatigue (FACIT-F) Scale

The FACIT-F scale (Cella and Webster 1997) is a brief, 13-item, symptom-specific questionnaire that specifically assesses the self-reported severity of fatigue and its impact upon daily activities and functioning. The FACIT-F questionnaire is presented below.



		Not at all	A little bit	Some-what	Quite a bit	Very much
HI7	I feel fatigued	0	1	2	3	4
HI 12	I feel weak all over	0	1	2	3	4
An 1	I feel listless ("washed out")	0	1	2	3	4
An 2	I feel tired	0	1	2	3	4
An 3	I have trouble starting things because I am tired	0	1	2	3	4
An 4	I have trouble finishing things because I am tired	0	1	2	3	4
An 5	I have energy	0	1	2	3	4
An 7	I am able to do my usual activities	0	1	2	3	4
An 8	I need to sleep during the day	0	1	2	3	4
An 12	I am too tired to eat	0	1	2	3	4
An 14	I need help doing my usual activities	0	1	2	3	4
An 15	I am frustrated by being too tired to do the things I want to do	0	1	2	3	4
An 16	I have to limit my social activity because I am tired	0	1	2	3	4

The FACIT-F uses a 13-item symptom-specific questionnaire to assess the burden of self-reported fatigue caused by a chronic disease and its impact upon daily activities and function.

A 5-point Likert-type scale (0 = not at all; 1 = a little bit; 2 = somewhat; 3 = quite a bit; 4 = very much) is used. As each of the 13 items of the FACIT-F scale ranges from 0‑4, the range of possible scores is 0‑52, with 0 being the worst possible score and 52 the best. To obtain the 0‑52 score, each negatively worded item response is transformed so that 0 is a bad response and 4 is good response. All responses are added with equal weight to obtain the total score. Lower values of the FACIT-F score denote higher fatigue.

Specifically, all questions with the exception of "I have energy" and "I am able to perform usual activities" will have their responses recoded so that 0 = Very much, 1 = Quite a Bit, 2 = Somewhat, 3 = A little bit, 4 = Not at all) prior to calculation of the FACIT-F scale score.

To score FACIT-Fatigue Subscale (FACIT_F), do the following:

1. Perform reversals for 11 out of the 13 individual items, namely HI7, HI12, An1, An2, An3, An4, An8, An12, An14, An15, and An16: item score = 4 minus item response
2. For the individual items An5 and An7: items score = item response
3. Sum the individual items to obtain a score
4. Multiply the sum of the items scores by 13 (the number of items in the subscale), then divide by the number of items answered which produces the Fatigue Subscale score.

Missing items are acceptable as long as more than 50% of the items are answered (i.e. a minimum of 7 out of 13 items). If less than 7 items are answered, the Fatigue Subscale score will be set to missing.

European Quality of Life-5 Dimensions-5 Level (EQ-5D-5L) Scores

The European Quality of Life-5 dimensions (EQ-5D) questionnaire is a widely used, generic questionnaire that assesses health status (The EuroQol Group 1990/Herdman et al. 2011). The questionnaire consists of 2 parts:

The first part assesses 5 dimensions: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. The levels of response for mobility, self-care, and usual activities are “no problems”, “slight problems”, “moderate problems”, “severe problems”, or “unable to”. For pain/discomfort, the levels of response are “no pain/discomfort”, “slight pain/discomfort”, “moderate pain/discomfort”, “severe pain/discomfort”, and “extreme pain/discomfort”. Lastly, for anxiety/depression, the levels of response are “not anxious or depressed”, “slightly anxious or depressed”, “moderately anxious or depressed”, “severely anxious or depressed”, and “extremely anxious or depressed”. This part of the EQ-5D can be used to generate a health state index score, which is often used to compute quality-adjusted life years (QALY) for utilization in health economic analyses. The health state index score is calculated based on the responses to the 5 dimensions, providing a single value on a scale from less than 0 (where zero is a health state equivalent to death; negative values are valued as worse than dead) to 1 (perfect health), with higher scores indicating better health utility.

The second part of the questionnaire consists of a visual analog scale (VAS) on which the patient rates their perceived health state from 0 (worst imaginable health state/the worst health you can imagine) to 100 (best imaginable health state/the best health you can imagine).

Health Assessment Questionnaire‑Disability Index (HAQ-DI)

The HAQ-DI is a patient-reported questionnaire that is commonly used in RA to measure disease-associated disability (assessment of physical function). It consists of 24 questions referring to 8 domains: dressing/grooming, arising, eating, walking, hygiene, reach, grip, and activities (Fries et al. 1980, 1982; Ramey et al. 1996).

1. Dressing and grooming (C1. Dress yourself, including tying shoelaces and doing buttons, C2. Shampooing your Hair)

   Includes 2 component questions, 1 device checkbox (devices used for dressing), 1 help checkbox.

2. Arising (C1. Stand up from straight chair, C2. Get in and out of bed)

   Includes 2 component questions, 1 device checkbox (built-up or special chair), 1 help checkbox.

3. Eating (C1. Cut your meat, C2. Lift a cup or glass to your mouth, C3. Open a new carton of milk)

   Includes 3 component questions, 1 device checkbox (built-up or special utensils), 1 help checkbox.

4. Walking (C1. Walk outdoors on flat ground, C2. Climb up five steps)

   Includes 2 component questions, 4 device checkboxes (cane, walker, crutches, wheelchair), 1 help checkbox.

5. Hygiene (C1. Wash and dry your body, C2. Take a tub bath, C3. Get on and off the toilet)

   Includes 3 component questions, 4 device checkboxes (raised toilet seat, bathtub seat, bathtub bar, long-handled appliances in bathroom), 1 help checkbox.

6. Reach (C1. Reach and get down a 5-pound object (such as a bag of sugar) from just above your head, C2. Bend down to pick up clothing from the floor)

   Includes 2 component questions, 1 device checkbox (long-handled appliances for reach), 1 help checkbox.

7. Grip (C1. Open car doors, C2. Open jars which have been previously opened, C3. Turn faucets on and off)

   Includes 3 component questions, 1 device checkbox (jar opener), 1 help checkbox.

8. Activities (C1. Run errands and shop, C2. Get in and out of a car, C3. Do chores such as vacuuming or yard work)

   Includes 3 component questions, 1 help checkbox.

In order to compute the HAQ-DI (Standard Disability Index) score, the following scores are assigned to the responses:

Without any difficulty = 0

With some difficulty = 1

With much difficulty = 2

Unable to do = 3.

The disability section of the questionnaire scores the patient's self-perception on the degree of difficulty (0 = without any difficulty, 1 = with some difficulty, 2 = with much difficulty, and 3 = unable to do) when dressing and grooming, arising, eating, walking, hygiene, reach, grip, and performing other daily activities. The reported use of special aids or devices and/or the need for assistance of another person to perform these activities is also assessed. The scores for each of the functional domains will be averaged to calculate the functional disability index.

Calculating the HAQ-DI:

The patient must have a score for at least 6 of the 8 categories. If there are less than 6 categories completed, a HAQ-DI cannot be computed.

- A category score is determined from the highest score of the sub-categories, or components, in that category. (For example, in the category EATING there are three sub-category items. If a patient responds with a 1, 2, and 0, respectively; the category score is 2.)

- Adjust for use of aids/devices and/or help from another person when indicated:
  - When there are no aids or devices or help indicated for a category, the category's score is not modified.
  - When aids or devices or help ARE indicated by the patient, adjust the score for a category by increasing a zero or a one to a two. If a patient's highest score for that sub-category is a two it remains a two, and if a three, it remains a three.
  - Sum the eight category scores
  - Divide the sum by the number of categories answered (range 6-8)

The scale is not truly continuous but has 25 possible values (i.e., 0, 0.125, 0.250, 0.375 ⋯ 3). The mapping of the aids or devices to the categories is the following:

HAQ-DI Category	Companion aids or devices item
Dressing and Grooming	Devices used for dressing (button hook, zipper pull, long handled shoe horn, etc.)
Arising	Built up or special chair
Eating	Built up or special utensils
Walking	Cane, walker, crutches
Hygiene	Long handled appliances in bathroom
Reach	Long handled appliances for reach
Grip	Jar opener (for jars previously opened)

Minimum Clinically Important Differences:

The summary and analyses of the HAQ-DI Score will include the number of patients achieving a MCID, which is defined as: MCID ≥0.22 reduction in HAQ-DI Score. Also, the number of patients having ≥0.3 reduction in HAQ-DI Score will be assessed.

Appendix 3. Corticosteroids

Corticosteroids

The following ATC code will be used to select all possible corticosteroids: H02 Corticosteroids for systemic use with route of Oral.

All unique preferred terms in the database falling under the above ATC code will be reviewed by the medical group in order to determine which ones should be included. All corticosteroid doses need to be converted to prednisone equivalent doses. If additional conversion factors are required, these will be added to the table below in a SAP prior to database lock.

The following table should be used for converting non-prednisone medications to prednisone equivalent:

Multiply the dose of the corticosteroid taken by the patient (in milligrams) in Column 1 by the conversion factor in Column 2 to get the equivalent dose of prednisone (in milligrams).

Example: Patient is taking 16 mg of methylprednisolone orally daily. To convert to prednisone: 16 mg methylprednisolone X 1.25 = 20 mg prednisone. 16 mg of methylprednisolone taken orally daily is equivalent to 20 mg of prednisone taken orally daily.

Column 1	Column 2
Corticosteroid Preferred Term	**Conversion factor for converting to an equivalent prednisone dose**
PREDNISONE	1
PREDNISONE ACETATE	1
PREDNISOLONE	1
PREDNISOLONE ACETATE	1
PREDNISOLONE SODIUM PHOSPHATE	1
METHYLPREDNISOLONE	1.25
METHYLPREDNISOLONE ACETATE	1.25
METHYLPREDNISOLONE SODIUM SUCCINATE	1.25
TRIAMCINOLONE	1.25
TRIAMCINOLONE ACETONIDE	1.25
TRIAMCINOLONE HEXACETONIDE	1.25
CORTISONE	0.2
CORTISONE ACETATE	0.2
HYDROCORTISONE	0.25
HYDROCORTISONE ACETATE	0.25
HYDROCORTISONE SODIUM SUCCINATE	0.25
BETAMETHASONE	6.25
BETAMETHASONE ACETATE	6.25
BETAMETHASONE DIPROPIONATE	6.25

BETAMETHASONE SODIUM PHOSPHATE	6.25
DEXAMETHASONE	6.25
DEXAMETHASONE ACETATE	6.25
DEXAMETHASONE PHOSPHATE	6.25
DEXAMETHASONE SODIUM PHOSPHATE	6.25
PARAMETHASONE	2.5
DEFLAZACORT	0.83
CELESTONA BIFAS	6.25
DEPO-MEDROL MED LIDOKAIN	1.25
DIPROSPAN	6.25
FLUOCORTOLONE	1
MEPREDNISONE	1.25

Appendix 4. X-ray Missing Data Imputation Plan

This document focuses on how to handle missing scores for individual joints. If after applying the following imputation method, the total score for modified Total Sharp Score (mTSS), erosion score (ER), and joint space narrowing (JSN) score are still missing, additional missing data imputation methods defined in Section 6.3.2 will be applied.

Sequence:

Two sets of scores will be generated by 2 independent readers at Bioclinica for all patients and provided to Lilly.

1. For each reader, identify missing joint scores and impute values where applicable
2. For each reader, identify any missing timepoints
3. If an adjudicated read is present, identify which reader pair will be used to derive scores
4. Derive mean scores from the reader pair selected

Rules:

- Missing individual joint scores for a given reader:
    - Missing joint score will be defined as joints scored as “S” (Surgically modified) or “NA” (not assessable due to poor radiographic depiction), as outlined in the Bioclinica imaging charter, or no score is provided for a joint because of missing x-ray.
    - For baseline:
        - If an individual joint for a patient has missing score at baseline, this joint will be set to be missing across all study visits and will not be used in calculation of mTSS, ER, and JSN total scores or change from baseline of these total scores during the entire study for this patient. If >50% of the joint scores at baseline are missing, the baseline will be set to missing and this patient will not be included in the analysis.
    - For postbaseline:
        - If >50% of the joint scores at a given timepoint are missing (including joints that have been set to missing as a result of missing baseline), this will be taken as an invalid/missing timepoint (visit), the mTSS, ER, and JSN total score will be set to missing for the timepoint. Additional imputation methods specified in Study SAP Section 6.3.2 will be applied to impute missing total score (see below for inter-reader discordance on missing timepoints).

        - If ≤50% of the joint scores at a given timepoint are missing and an individual joint score becomes missing postbaseline, the score for this joint will be imputed based on an LOCF approach, using the last previous timepoints (including baseline) for which a non-missing score is available for the joint.

- Missing timepoints
    - If a time point is set to invalid/missing based on both independent readers, this timepoint will be taken as a missing timepoint.
    - If adjudication occurs because a timepoint was set as missing for only one of the two initial independent readers, then the reader pair selected and handling of the timepoint will depend on whether the adjudication reader's score for the timepoint concerned is set as missing or not missing:
        - If the adjudication reader's score for the timepoint concerned is not set as missing, the pair used will include the adjudication reader and the other independent reader for whom that timepoint was not set as missing. This will not be handled as a missing timepoint.
        - If the adjudication reader's score for the timepoint concerned is set as missing, the pair will include the adjudication reader and the other independent reader for whom that timepoint was set as missing. This will be handled as a missing timepoint and the mTSS, ER, and JSN total score will be set to missing. Additional imputation methods specified in Study SAP Section 6.3.2 will be applied to impute missing total score.
- General approach to scoring
    - For description of static scores at a given timepoint, the mean of the scores generated by the two independent readers at that timepoint (following imputation of missing scores if applicable) will be used.
    - For calculation of changes in mTSS, ER, and JSN score, the change from baseline will first be calculated for the 2 independent readers separately (following imputation of missing scores if applicable). The change score will then be taken as the mean of these 2 independent change scores.
    - In the case of adjudicated patients, a third set of scores generated by an independent adjudication reader will be provided (as outlined in the Bioclinica Imaging Charter).
        - Where an adjudication reader is used, Lilly will use scores generated by the adjudication reader and the other independent reader whose score for change from baseline in mTSS to Week 24 (the primary time point for structure data analysis) is closest to that generated by the adjudication

reader (following imputation of missing scores if applicable) as a general principle.

- If the smallest difference at Week 24 cannot be determined, then move back to the previous timepoint to determine the pair.

Footnotes:

a. In such cases, the last valid (non-imputed) score for that individual joint will be used for imputation, even if it is obtained from a timepoint set as missing as a result of >50% missing joints.

Leo Document ID = 0285312d-040c-4bca-8d18-53100ab3d8ce

Approver: PPD
Approval Date & Time: 12-Jan-2017 03:04:32 GMT
Signature meaning: Approved