1.0 Title Page

Statistical Analysis Plan

Study M15-572

A Phase 3, Randomized, Double-Blind, Study Comparing Upadacitinib (ABT-494) to Placebo and to Adalimumab in Subjects with Active Psoriatic Arthritis Who Have a History of Inadequate Response to at Least One Non-Biologic Disease Modifying Anti-Rheumatic Drug (DMARD) – *SELECT – PsA 1*

Date: 08 Jul 2021

Version 4.0

2.0 Table of Contents

List of Tables

List of Figures

List of Appendices

3.0 Introduction

This statistical analysis plan (SAP) describes the statistical analyses to be completed by the Data and Statistical Science Department for upadacitinib Study M15-572. It provides details to further elaborate statistical methods as outlined in the protocol. Pharmacokinetic and biomarker analyses will be performed separately and the corresponding analysis plan is documented separately are not in the scope of this SAP.

Unless noted otherwise, all analyses will be performed using SAS version 9.2 or later (SAS Institute Inc., Cary, NC 27513) under the UNIX operating system.

4.0 Study Objectives, Design and Procedures

4.1 Study Objectives

Period 1

Primary Objective

1. To compare the efficacy of upadacitinib 15 mg once daily (QD) and 30 mg QD versus placebo and versus adalimumab (ADA) 40 mg every other week (eow) for the treatment of signs and symptoms in subjects with moderately to severely active Psoriatic Arthritis (PsA) who have an inadequate response or intolerance to 1 or more non-biologic DMARD (DMARD-IR).

Secondary Objective

2. To compare the efficacy of upadacitinib 15 mg QD and 30 mg QD versus placebo for the prevention of structural progression in subjects with moderately to severely active PsA who have an inadequate response or intolerance to 1 or more non-biologic DMARD (DMARD-IR).
3. To compare the safety and tolerability of upadacitinib 15 mg QD and 30 mg QD versus placebo and versus adalimumab in subjects with moderately to severely

active PsA who have an inadequate response or intolerance to 1 or more non-biologic DMARD (DMARD-IR).

Period 2

To evaluate the long-term safety, tolerability and efficacy of upadacitinib 15 mg QD and 30 mg QD in subjects with PsA who have completed Period 1.

4.2 Overall Study Design and Plan

This is a Phase 3 multicenter study that includes two periods. Period 1 is 56 weeks in duration and includes a 24-week randomized, double-blind, parallel-group, placebo-controlled and active-comparator controlled period followed by an additional 32 weeks of blinded active comparator-controlled treatment (Weeks 24 – 56). Period 1 is designed to compare the safety, tolerability, and efficacy of upadacitinib 15 mg QD and 30 mg QD versus placebo and versus adalimumab 40 mg every other week (eow) in subjects with moderately to severely active PsA who have an inadequate response to non-biologic DMARDs (DMARD-IR). Period 1 is also designed to compare the efficacy of upadacitinib 15 mg QD and 30 mg QD versus placebo for the prevention of structural progression. Period 2 is an open-label (blinded until the last subject completes the last visit of Period 1), long-term extension of up to a total treatment duration of approximately 5 years to evaluate the safety, tolerability and efficacy of upadacitinib 15 mg QD and 30 mg QD in subjects with PsA who have completed Period 1. Starting with Amendment 6.01, subjects in Japan receiving upadacitinib 30 mg QD will be switched to upadacitinib 15 mg QD. For other countries, starting with Amendment 7 and 7.01, subjects receiving upadacitinib 30 mg QD will be switched to upadacitinib 15 mg QD.

The study is designed to enroll approximately 1640 subjects at approximately 350 study centers worldwide to meet scientific and regulatory objectives without enrolling an undue number of subjects in alignment with ethical considerations.

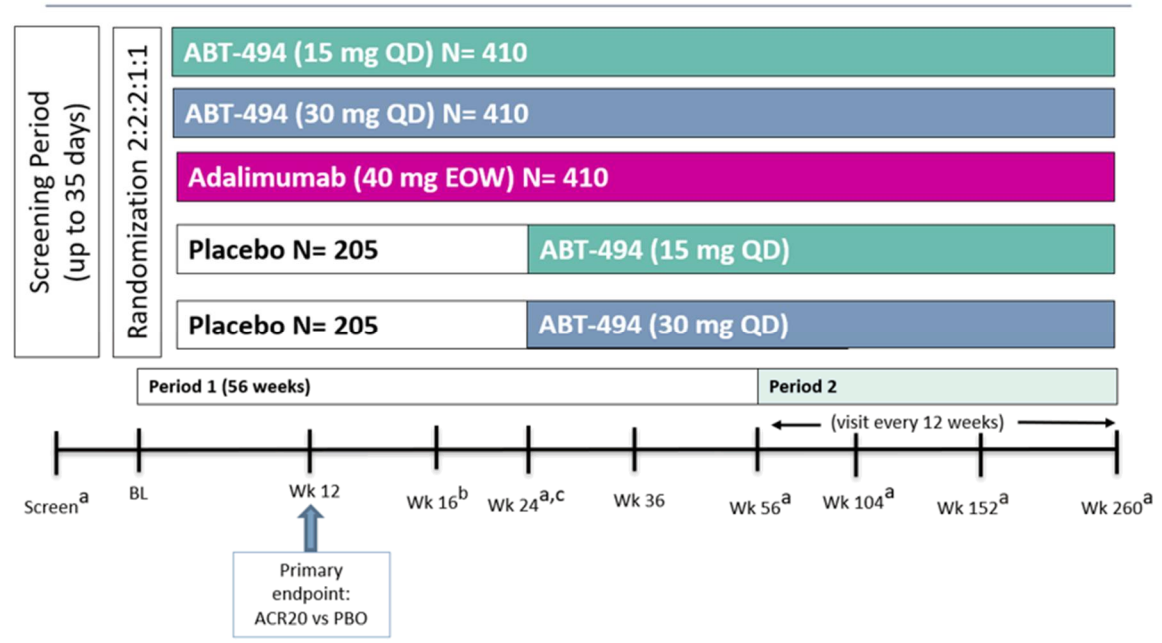
The study duration will include a 35-day screening period; a 56-week blinded period which includes 24 weeks of double-blind, placebo-controlled and active comparator-controlled treatment followed by 32 weeks of active comparator-controlled treatment (Period 1); a long-term extension period of up to a total treatment duration of approximately 5 years ([blinded to sites and investigators until the last subject completes the last visit of Period 1] Period 2); and a 30-day follow-up call or visit; and a 70-day follow-up call. Subjects who meet eligibility criteria will be randomized in a 2:2:2:1:1 ratio using an Interactive Response Technology (IRT) to receive double-blind study drug in one of the following treatment groups:

- Group 1: Upadacitinib 15 mg QD, N = 410 (Day 1 to Week 24) → Upadacitinib 15 mg QD (Week 24 and thereafter)
- Group 2: Upadacitinib 30 mg QD, N = 410 (Day 1 to Week 24) → Upadacitinib 30 mg QD (Week 24 and thereafter)
- Group 3: Adalimumab (40 mg eow) (N = 410)
- Group 4: Placebo, N = 205 (Day 1 to Week 24) → Upadacitinib 15 mg QD (Week 24 and thereafter)
- Group 5: Placebo, N = 205 (Day 1 to Week 24) → Upadacitinib 30 mg QD (Week 24 and thereafter)

Randomization will be stratified by extent of psoriasis (≥ 3% body surface area [BSA] or < 3% BSA), current use of at least 1 DMARD (Yes or No), presence of dactylitis, and presence of enthesitis, except for subjects from China and Japan, for which randomization will be stratified by extent of psoriasis (≥ 3% body surface area [BSA] or < 3% BSA) only. See Section 5.5.3 of protocol for details.

A schematic of the overall study design is shown in Figure 1 below.

Figure 1. Study Design



a. All subjects will receive x-rays of hands and feet at Screening, Wk 24, Wk 56, Wk 104, Wk152, and Wk 260/PD.

b. At Week 16 rescue therapy will be offered to subjects classified as non-responders (defined as not achieving at least 20% improvement in either or both tender joint count (TJC) and swollen joint count (SJC) at both Week 12 and Week 16) as described in Protocol Section 5.2.3.4.

c. At Week 24, all placebo subjects will switch to Upadacitinib 15 mg QD or 30 mg QD (1:1 ratio) regardless of response.

Note: For subjects in Japan only: Upon approval of protocol amendment 6.01, subjects receiving upadacitinib 30 mg QD will be switched to upadacitinib 15 mg QD prior to marketing approval in Japan for the treatment of psoriatic arthritis.

Note: Upon approval of protocol amendment 7 and 7.01, subjects receiving upadacitinib 30 mg QD will be switched to upadacitinib 15 mg QD at their next scheduled study visit in other countries.

4.3 Sample Size

The planned total sample size of approximately 1640 for this study (with 2:2:2:1:1 randomization ratio for upadacitinib 15 mg, upadacitinib 30 mg, adalimumab 40 mg eow and placebo subjects) provides at least 90% power for a 20% difference in ACR20 response rate (assuming a placebo ACR20 response rate of 30%). It will also

provide at least 90% power for the majority of the key secondary endpoints. With the given sample size, there is approximately 90% power to detect a standardized effect size of 0.26 in change from baseline in SHS for each upadacitinib dose group versus the combined placebo group at Week 24. This sample size will also provide at least 85% power for evaluating non-inferiority for each upadacitinib dose group vs. adalimumab in ACR20 response rate at Week 12 assuming 50% ACR20 response rates for adalimumab and upadacitinib and 30% ACR20 response rates for placebo. All power and sample size calculations are performed at a two-sided significance level of 0.025 and accounting for a 10% dropout rate.

4.4 Week 24 Analysis and Data Base Lock

After the last subject completes the Week 24 study visit, an unblinded analysis will be conducted for the purpose of regulatory submission. To maintain integrity of the trial during the blinded 56-week period (Period 1), study sites and subjects will remain blinded until all subjects have reached Week 56.

4.5 Data Monitoring Committee (DMC) Activities

An independent external Data Monitoring Committee (DMC) is used to review unblinded safety data at regular intervals during the conduct of the study. The DMC will provide recommendation to an AbbVie Point of Contact on whether to continue, modify, or terminate studies after each review. When needed, high-level unblinded efficacy data may also be requested by the DMC and be reviewed so that the DMC can assess benefit:risk of any emerging safety differences.

An unblinded interim efficacy analysis will also be conducted by an independent external DMC after at least 600 subjects have completed the Week 12 Visit or have prematurely discontinued from the study. The interim analysis is to assess if the study met the pre-defined No-Go boundary for the primary endpoint ACR20. A futility recommendation will be made only when both the upadacitinib 15 mg and 30 mg QD doses meet the futility criteria. To address potential practical regulatory and statistical concerns, a small alpha of 0.0001 will be spent for the interim analysis and the final analysis will be

performed at level 0.0499. The overall Type I error rate is controlled at 0.05. However, in no circumstance should a decision be made to stop the trial early for superiority in efficacy of ABT-494 versus either adalimumab or placebo. Details of the interim efficacy analysis are described in the DMC SAP and details of the decision making rules are specified in the DMC charter.

5.0 Analysis Populations and Analysis Windows

5.1 Analysis Populations

Full Analysis Set (FAS)

The Full Analysis Set (FAS) includes all randomized subjects who received at least one dose of study drug. The FAS will be used for all efficacy and baseline analyses.

Per Protocol Analysis Set

The Per Protocol Analysis Set represents a subset of the FAS and consists of all FAS subjects who did not meet any major protocol deviations that are determined to have a potential impact on the primary efficacy endpoint up to Week 12 in Period 1 of the study. Additional analysis of the primary efficacy endpoint will be conducted on the Per Protocol analysis set, in order to evaluate the impact of major protocol deviations.

Protocol deviations with potential to affect the primary efficacy endpoint (key ICH deviations and other clinically significant non-ICH deviations) will be identified prior to the Week 24 database lock.

Safety Analysis Set

The Safety Analysis Set consists of all subjects who received at least one dose of study drug. For the Safety Analysis Set, subjects will be analyzed "as treated," regardless of the treatment randomized. "As treated" is determined by the treatment the subject received during the majority of the subject's drug exposure time in the analysis period.

5.2 Analysis Windows

Definition of Study Days (Days Relative to the First Dose of Study Drug)

Study Days are calculated for each collection date relative to the date of the first dose of study drug. It is defined as the number of days between the date of the first dose of study drug and the collection date. Study days are negative values when the collection date of interest is prior to the first study drug dose date. Study days are positive values when the collection date of interest is on or after the first study drug dose date. The day of the first dose of study drug is defined as Study Day 1, while the day prior to the first study drug dose is defined as Study Day –1 (there is no Study Day 0). Study days are used to map actual study visits to the protocol-specified study visits.

Definition of Analysis Windows

The following rules will be applied to assign actual subject visits to protocol-specified visits. For each protocol-specified study visit, a target study day will be identified to represent the corresponding visit along with a window around the target day. Windows will be selected in a non-overlapping fashion so that a collection date does not fall into multiple visit windows. If a subject has two or more actual visits in one visit window, the visit closest to the target day will be used for analysis. If two visits are equidistant from the target day, then the later visit will be used for analysis.

The visit window and the target study day for each protocol-specified visit in Period 1 are displayed in Table 1, Table 2, Table 3, Table 4, Table 5 and Table 6 (depending on the different visit schedules of different endpoints). Visit windows for protocol-specified visits in Period 2 are defined similarly.

Table 1. Analysis Windows for Efficacy Analysis for Period 1 (for ACR Components, TJC28, SJC28, ESR and Derived Endpoints[a]) and Safety Analysis for Period 1 (for Lab Values [Clinical Chemistry, Hematology, Urinalysis] and Vital Signs/Weight)

Protocol Specified Visit Week	Lower Bound	Target Day	Upper Bound
Baseline	–99	1[b]	1
2	2	15	22
4	23	29	43
8	44	57	71
12	72	85	99
16	100	113	127
20	128	141	155
24	156	169	Min (183, first dose date from Week 24 dispensed kit)
28	Min (183, first dose date from Week 24 dispensed kit) + 1	197	211
32	212	225	239
36	240	253	281
44	282	309	351
56	352	393	435

a. ACR20/50/70 response rates, DAS28 (CRP), DAS28 (ESR), PsARC, DAPSA, proportion of subjects achieving a clinically meaningful improvement in HAQ-DI (≥ 0.35).

b. Day of first dose of study drug.

Table 2. Analysis Windows for Efficacy Analysis for Period 1 (for SF-36, EQ-5D-5L, FACIT-F, WPAI, BASDAI, HRU and Derived Endpoints[a])

Protocol Specified Visit Week	Lower Bound	Target Day	Upper Bound
Baseline	–99	1[b]	1
12	2	85	127
24	128	169	Min (211, first dose date from Week 24 dispensed kit)
36	Min (211, first dose date from Week 24 dispensed kit) + 1	253	323
56	324	393	435

a. PASDAS, BASDAI 50 response rates, Change from baseline in Morning stiffness (mean of BASDAI Questions 5 and 6), ASDAS, proportion of subjects with ASDAS Inactive Disease, proportion of subjects with ASDAS Major Improvement, proportion of subjects with ASDAS Clinically Important Improvement.

b. Day of first dose of study drug.

Table 3. Analysis Windows for Safety Analysis for Period 1 (for Total Cholesterol, HDL-C, LDL-C, Triglycerides)

Protocol Specified Visit Week	Lower Bound	Target Day	Upper Bound
Baseline	–99	1[a]	1
4	2	29	57
12	58	85	127
24	128	169	Min (211, first dose date from Week 24 dispensed kit)

a. Day of first dose of study drug.

Table 4. Analysis Windows for Efficacy Analysis for Period 1 (for BSA-Psoriasis, PASI, sIGA, LDI, LEI, SPARCC Enthesitis Index and Derived Endpoints[a])

Protocol Specified Visit Week	Lower Bound	Target Day	Upper Bound
Baseline	–99	1[b]	1
12	2	85	99
16	100	113	141
24	142	169	Min (211, first dose date from Week 24 dispensed kit)
36	Min (211, first dose date from Week 24 dispensed kit) + 1	253	323
56	324	393	435

a. MDA, dactylitis count/total enthesitis count, proportion of subjects with resolution of dactylitis/enthesitis (out of all sites, or the sites included in SPARCC or LEI), PAISI75/90/100, proportion of subjects achieving a static Investigator Global Assessment of psoriasis (sIGA) score of 0 or 1 and at least a 2-point improvement from baseline.

b. Day of first dose of study drug.

Table 5. Analysis Windows for Efficacy Analysis for Period 1 (for SAPS)

Protocol Specified Visit Week	Lower Bound	Target Day	Upper Bound
Baseline	–99	1[a]	1
16	2	113	141
24	142	169	Min (211, first dose date from Week 24 dispensed kit)
36	Min (211, first dose date from Week 24 dispensed kit) + 1	253	323
56	324	393	435

a. Day of first dose of study drug.

Table 6. Analysis Windows for Efficacy Analysis for Period 1 (for X-Rays of Hands and Feet)

Protocol Specified Visit Week	Lower Bound	Target Day	Upper Bound
Baseline	–99	1[a]	1
16[b]	2	113	141
24	142	169	197
36[c]	198	253	323
56	324	393	435

a. Day of first dose of study drug.

b. Week 16 x-rays will be performed for all subjects who met the protocol specified rescue criteria at Week 16 (i.e., not achieving at least 20% improvement in either or both TJC68 and SJC66 at both Week 12 and Week 16).

c. Not protocol specified visit for x-ray assessment, but will be used to capture unscheduled or PD x-rays.

6.0 Demographics, Baseline Characteristics, Medical History, and Previous/Concomitant Medications

6.1 Demographics and Baseline Characteristics

Demographic and baseline characteristics information will be collected at the Baseline visit of the study and will be summarized for the FAS. The number of observations, mean, standard deviation, median, minimum and maximum will be summarized for continuous variables. Categorical or discrete variables will be summarized via frequencies and percentages. Summary statistics will be computed for each treatment group and overall.

Main Demographic and Baseline Characteristics

- Sex (male, female)
- Age (years)
- Age Categories (< 65, ≥ 65 – 75, ≥ 75 years)
- Race (White, Black or African American, American Indian or Alaska Native, Native Hawaiian or Other Pacific Islander, Asian, Other)

- Geographic Region (North America, Western Europe and Oceania, Eastern Europe, Latin-America, Asia, and Other)
- Ethnicity (Hispanic or Latino, Not Hispanic or Latino)
- Weight (kg)
- Height (cm)
- Body Mass Index (BMI) (kg/m^2)
- Body Mass Index (BMI) Category (kg/m^2) (BMI < 25, BMI ≥ 25)

PsA Medical History, Prior and Baseline Treatments

- Duration of PsA symptoms in years
- Duration of PsA diagnosis in years
- Rheumatoid Factor (RF) status: Positive or Negative
- Anti-CCP status: Positive or Negative
- Number of prior non-biologic DMARDs (0, 1, 2, ≥ 3)
- Current use of at least 1 non-biologic DMARD at baseline (Yes or No)
- Concomitant non-biologic DMARD at baseline (MTX alone, MTX and other non-biologic DMARD, non-biologic DMARD other than MTX)
- Current NSAID use at baseline (Yes or No)
- Current Corticosteroid use at baseline (Yes or No)

Baseline Disease Characteristics

- Tender joint count (TJC68) defined as the number of tender joints out of 68 assessed joints
- Swollen joint count (SJC66) defined as the number of swollen joints out of 66 assessed joints
- Physician's global assessment of Disease Activity Numeric Rating Scale (0 - 10 NRS)
- Patient's assessment of pain (0 - 10 NRS)
- Patient's global assessment of disease activity (0 - 10 NRS)

- Health Assessment Questionnaire Disability Index (HAQ-DI)
- High sensitivity C-reactive protein (hsCRP)
- Categories for hsCRP: equal or below vs. above ULN (2.87 mg/L)
- Body Surface Area (BSA) with Psoriasis (as categorical with < 3%, ≥ 3%)
- Body Surface Area (BSA) with Psoriasis (as continuous for subjects with BSA-Ps > 0%)
- Static Investigator Global Assessment of Psoriasis (sIGA) (as a categorical variable)
- Psoriasis Area and Severity Index (PASI) score (for subjects with ≥ 3% Body Surface Area (BSA) psoriasis)
- Erythrocyte sedimentation rate (ESR)
- Psoriatic arthritis disease activity score (PASDAS)
- Disease Activity In Psoriatic Arthritis (DAPSA) score
- Presence of Dactylitis (LDI > 0)
- Leeds Dactylitis Index (LDI) (out of subjects with presence of dactylitis (LDI > 0))
- Total dactylitis count (out of subjects with presence of dactylitis (LDI > 0))
- Presence of enthesitis (total enthesitis count > 0)
- Total enthesitis count (out of subjects with presence of enthesitis (total enthesitis count > 0))
- Spondyloarthritis Research Consortium of Canada (SPARCC) Enthesitis Index (out of subjects with SPARCC Enthesitis Index > 0)
- Leeds Enthesitis Index (LEI) (out of subjects with LEI > 0)
- DAS28 [CRP]
- DAS28 [ESR]
- Presence of Psoriatic Spondylitis
- Ankylosing Spondylitis Disease Activity Score (ASDAS) (out of subjects with presence of psoriatic spondylitis)
- Bath AS Disease Activity Index (BASDAI) (out of subjects with presence of psoriatic spondylitis)

- Morning stiffness (mean of BASDAI Questions 5 and 6)
- Morning stiffness severity 0 - 10 (BASDAI Question 5)
- Morning stiffness duration 0 - 10 (BASDAI Question 6)
- Radiographic endpoints:
  - Presence of ≥ 1 erosion
  - Total modified PsA Sharp/van der Heijde Score (SHS) score
  - Erosion score
  - Joint space narrowing score

Patient Report Outcomes at Baseline

- Functional Assessment of Chronic Illness Therapy (FACIT)-Fatigue
- 36-Item Short Form Health Survey (SF-36) Version 2: physical component summary, mental component summary, and the 8 sub-domain scores
- EuroQol-5D-5L (EQ-5D-5L) index and VAS score
- Work Productivity and Activity Impairment (WPAI)
- Self-Assessment of Psoriasis Symptoms (SAPS)

Tobacco/Nicotine and Alcohol Use

- Tobacco/Nicotine Use [user, ex-user, non-user, unknown]
- Alcohol Use [drinker, ex-drinker, non-drinker, unknown]

6.2 Medical History

Medical history data will be summarized and presented for FAS population using System Organ Class (SOC) and Preferred Term (PT) as in the Medical Dictionary for Regulatory Activities (MedDRA) dictionary. The number and percentage of subjects with a particular PT will be summarized for each randomized treatment group as well as overall. Subjects reporting more than one PT within a SOC will be counted only once for that SOC. No statistical comparison will be performed for medical history reporting.

6.3 Prior Treatment and Concomitant Medications

Prior and concomitant medications will be summarized by each randomized treatment group as well as overall for FAS. Prior medications are those medications taken prior to the first dose of study drug. This includes medications with a start date before the first study drug administration date, regardless of the end date of these medications. Medications taken on the day of the first dose of study drug are not counted as prior medications. Concomitant medications are those medications, other than study drug, taken after the first dose of study drug and within 1 day of the last dose of study drug for the subjects on upadacitinib and placebo, and within 14 days of the last dose of study drug for the subjects on adalimumab, respectively. This includes medications with a start date between first study drug administration and last study drug administration + 1 or 14 day, as well as, medications with a start date prior to first dose of study drug and which are ongoing after first dose of study drug. Medications taken on the day of the first dose of study drug are counted as concomitant medications.

The number and percentage of subjects who received a prior medication and the number and percentage of subjects who received a concomitant medication will be tabulated separately by the generic name assigned by the most current version of the World Health Organization (WHO) Drug Dictionary.

6.4 Protocol Deviations

Protocol deviations based on ICH deviation criteria are categorized as follows:

1. Those who entered the study even though they did not satisfy the entry criteria
2. Those who developed withdrawal criteria during the study and were not withdrawn
3. Those who received the wrong treatment or incorrect dose, and
4. Those who received an excluded or prohibited concomitant medication.

The protocol deviations listed above will be summarized and listed by treatment group.

7.0 Patient Disposition

The following will be summarized by randomized treatment group as well as overall:

- number of subjects randomized,
- number of subjects included in key analysis populations (Full Analysis Set and Per Protocol Analysis Set for efficacy analysis, Safety Analysis Set for safety analysis),
- number of subjects who completed Period 1 study participation,
- number of subjects who entered Period 2,
- number of subjects who completed overall study (Period 1 and Period 2) participation (if applicable).

Premature discontinuation details will be further summarized separately for Period 1 and Period 2 as follows.

Period 1

The number and percentage of subjects completed Period 1 and prematurely discontinued in Period 1 will be summarized by randomized treatment group, separately by study drug and study participation completion/discontinuation, with the reason for discontinuation collected from CRF by the following categories:

- Adverse event (AE)
- Withdrew consent
- Lost to follow-up
- Lack of efficacy
- Other.

In addition, the number and percentage of subjects who entered Period 2 will also be summarized by randomized treatment group.

For Week 24 reporting, this summary will also be presented for study drug completion/discontinuation by Week 24.

Period 2

Period 2 patient dispositions and reason for discontinuation will be summarized with the same categories as given above for Period 1 for overall total and by treatment in Period 2 defined as follows:

1. Upadacitinib 15 mg QD
2. Upadacitinib 30 mg QD
3. Adalimumab 40 mg EOW

Among the subjects who entered Period 2 participation (regardless of whether subject entered Period 2 on study drug), the number and percentage of subjects who completed and prematurely discontinued study participation in Period 2 will be summarized.

Among the subjects who entered Period 2 on study drug, the number and percentage of subjects who completed and prematurely discontinued study drug in Period 2 will be summarized.

Summary of Dose Switch from Upadacitinib 30 mg QD to 15 mg QD

Starting with Amendment 6.01 (for Japan) and Amendment 7 and 7.01(for global), subjects receiving upadacitinib 30 mg QD will be switched to upadacitinib 15 mg QD. The visit at which dose switch occurs could be different for each subject. For subjects on upadacitinib 30 mg QD in Period 2, the number and percentage of subjects switching to upadacitinib 15 mg QD at each visit will be summarized.

8.0 Study Drug Exposure and Compliance

8.1 Study Drug Exposure

For short term up to Week 24, the duration of exposure to study drug will be summarized for the safety analysis set by the randomized treatment groups.

For long term, the duration of exposure to study drug will be summarized for the safety analysis set by the following groups.

1. Upadacitinib 15 mg QD

   This includes upadacitinib 15 mg QD exposure from subjects starting on upadacitinib 15 mg QD and subjects switching from placebo to upadacitinib 15 mg QD.

2. Upadacitinib 30 mg QD

   This includes upadacitinib 30 mg QD exposure from subjects starting on upadacitinib 30 mg QD and subjects switching from placebo to upadacitinib 30 mg QD. Exposure is censored at time of dose switch from upadacitinib 30 mg QD to upadacitinib 15 mg QD.

3. Upadacitinib 15 mg QD switched from upadacitinib 30mg QD

   This includes Upadacitinib 15 mg QD exposure from subjects who switched dose from upadacitinib 30 mg QD to upadacitinib 15 mg QD.

4. Adalimumab 40 mg EOW

Exposure to upadacitinib and placebo is defined as last dose date minus first dose date plus 1 day. Exposure to adalimumab is defined as last dose date minus first dose date plus 14 days.

Starting with Amendment 6.01 (for Japan) and Amendment 7 and 7.01(for global), subjects receiving upadacitinib 30 mg QD will be switched to receiving upadacitinib 15

mg QD. Exposure to upadacitinib 30 mg QD (Groups 2) will be censored prior to the day subject received the first dose of upadacitinib 15 mg QD, and subsequent exposure to upadacitinib 15 mg QD will be summarized under separate groups (Groups 3).

The duration of exposure to study drug will be summarized for each group as specified above, with the number of subjects, mean, standard deviation, median, minimum and maximum values. In addition, the number and percentage of subjects exposed to study drug will be summarized for the following cumulative duration intervals.

- ≥ 2 weeks
- ≥ 1 month
- ≥ 3 months
- ≥ 6 months
- ≥ 9 months
- ≥ 12 months
- ≥ 18 months
- ≥ 2 years
- ≥ 2.5 years
- ≥ 3 years
- ≥ 3.5 years
- ≥ 4 years
- ≥ 4.5 years

8.2 Compliance

Study drug compliance will be summarized for each treatment group at Week 24. The summary will be performed for UPA/PBO tablets and ADA/PBO injections separately. For tablets, compliance is defined as the number of tablets taken (i.e., the difference between the number of tablets dispensed and the number of tablets returned) during the subject's participation up to Week 24 divided by the number of days that the subject was in the Treatment Phase up to Week 24. For injections, compliance is defined as the

number of ADA injections administered during the subject's participation up to Week 24 divided by the number of injections planned during the subject's participation in the Treatment Phase up to Week 24.

9.0 Efficacy Analysis

9.1 General Considerations

There are two sets of planned efficacy analysis: efficacy analysis through Week 24 and long-term efficacy analysis. Unless otherwise noted, all efficacy analyses will be carried out using the FAS population.

9.1.1 Efficacy Analysis at Different Phases of the Study

Efficacy Analysis by Week 24

Standard efficacy analysis by randomized treatment groups (upadacitinib 15 mg QD, upadacitinib 30 mg QD, adalimumab 40 mg EOW and the combined placebo groups) will be performed on efficacy data up to Week 24. No protocol-defined treatment switching will occur prior to the time point. Formal statistical inference will be generated, and results from this set of analysis will be used as the key efficacy findings of this study.

Long-Term Efficacy Analysis

Long-term efficacy analysis will be performed on As Observed data (defined in Section 9.1.2) by randomized treatment group sequence as described below:

1. Placebo → Upadacitinib 15 mg QD
2. Placebo → Upadacitinib 30 mg QD
3. Upadacitinib 15 mg QD
4. Upadacitinib 30 mg QD
5. Adalimumab 40 mg EOW

Descriptive statistics and 95% confidence intervals will be provided for all treatment sequences. In addition, comparisons between each upadacitinib dose and adalimumab (treatment sequences 3, 4, and 5) will be provided up to Week 56. Treatment comparisons may also be provided after Week 56 as appropriate.

For long term efficacy analysis, subjects will continue to be summarized under the treatment sequences as described above, regardless of dose switch from upadacitinib 30 mg QD to upadacitinib 15 mg QD – i.e., for treatment sequences 2 and 4, data collected after dose switch will continue to be summarized under the same treatment sequences. The visit at which dose switch occurs could be different for each subject. The first and last visits at which dose switch occurs will be noted in the summary. In addition, sensitivity analyses will be performed where the subjects who switched from upadacitinib 30 mg QD to 15 mg QD will be censored at the time of dose switch, and efficacy values before and post switch will be summarized separately.

9.1.2 Handling of Missing Data and Intercurrent Events

Non-Responder Imputation (NRI) Approach

The NRI approach will handle data for binary endpoints as follows.

- Subjects who prematurely discontinue from study drug will be considered as non-responders for all subsequent visits after discontinuation.
- In addition, any subject with any missing value for the binary endpoints at a specific visit will be treated as non-responder for that visit.

NRI data handling will be used for the primary estimand (refer to Section 9.2.1) for the binary endpoints.

Mixed-Effects Model Repeated Measures (MMRM) Up to Week 24

The repeated measure analysis will be conducted using mixed models including observed data at all visits. For the MMRM analysis, data collected after premature discontinuation

of study drug will be excluded. The mixed model includes the categorical fixed effects of treatment, visit, treatment-by-visit interaction, the stratification factor current DMARD use (yes/no) and the continuous fixed covariates of baseline measurement. An unstructured variance covariance matrix will be used. The parameter estimations are based on the assumption of data being missing at random (MAR) and using the method of restrictive maximum likelihood (REML). MMRM will be used for the primary estimand of the non-radiographic continuous endpoints (refer to Section 9.2.3).

Linear Extrapolation for Radiographic Data

For radiographic data (i.e., SHS-based endpoints), linear extrapolation will be applied to subjects with missing data including those due to premature discontinuation of study drug or meeting the rescue criteria, where the missing x-ray at the time point of interest will be imputed assuming a linear relationship across visits. Linear extrapolation will be used to facilitate the primary estimand for radiographic endpoints (refer to Section 9.2.3). Linear extrapolation will not be provided for Week 104 analysis and future analyses.

As Observed (AO)

The AO data handling will not impute values for missing evaluations, and thus a subject who does not have an evaluation on a scheduled visit will be excluded from the AO analysis for that visit. Regardless of premature discontinuation of study drug or use of rescue medication, all observed data will be used in the analysis. The AO data handling will be used to facilitate the supplementary analysis for both binary and continuous endpoints including radiographic endpoints (refer to Section 9.2.2 and Section 9.2.3 where the corresponding supplementary estimands for the primary and ranked secondary endpoints are described respectively).

The intention-to-treat (ITT) principle for the analysis using AO data, AO (with imputation), where additional missing data will be imputed as non-responders, will be used for the supplementary analysis of the primary and key secondary binary endpoints. AO with imputation will not be provided for Week 104 analysis and future analyses.

Mixed Effect Model Repeat Measurement (MMRM) and Generalized Linear Mixed Model (GLMM) for Long-Term Analysis

The repeated measure analysis will be conducted using mixed model including As Observed measurements at all visits. MMRM will be used for continuous endpoints and GLMM will be used for non-radiographic binary endpoints. The mixed models will include the categorical fixed effects of treatment, visit and treatment-by-visit interaction, and stratification factor current DMARD use (yes/no). The categorical fixed effect of subject's discontinuation status may also be included in the model as appropriate. For the MMRM analysis of change from baseline in continuous endpoints, the baseline measurement will be included as a continuous fixed covariate. Unstructured, Toeplitz, compound symmetry, or other covariance structures may be considered.

9.2 Efficacy Analysis through Week 24

9.2.1 Primary Efficacy Analysis of the Primary Efficacy Endpoint

The primary efficacy endpoint is ACR20 response at Week 12. The primary estimand is the difference in the proportion of PsA patients who achieved ACR20 response at Week 12 and did not discontinue study drug by Week 12, comparing those who are randomized to each upadacitinib dose group and received study drug to those who are randomized to placebo and received study drug.

Analysis of the primary endpoint will be conducted on the FAS based on randomized treatment groups (upadacitinib 15 mg QD, upadacitinib 30 mg QD versus the combined placebo groups). Point estimate and 95% CI using normal approximation will be provided for the response rate for each randomized treatment group. Comparisons of the primary endpoint will be made between each upadacitinib dose and the combined placebo group using the Cochran-Mantel-Haenszel test (CMH) (refer to Appendix A) adjusting for the main stratification factor of current DMARD use (yes/no). Point estimate, 95% CI using normal approximation and p-value for the treatment comparison will be presented. Nominal p-values constructed using the CMH test will be provided. The multiplicity adjusted (as described in Section 9.2.5) testing results (significant or not significant) will

also be provided. NRI data handling will be used to facilitate the primary estimand. In addition, the number and percentage of non-responders by the intercurrent events, such as discontinuation of study drug prior to or at Week 12 or missing Week 12 ACR measurement will be summarized.

9.2.2 Supplementary Analysis of the Primary Efficacy Endpoint

For the primary efficacy endpoint, the same CMH analysis as detailed in Section 9.2.1 will be repeated using As Observed (AO) data handling without any imputation as supplementary analysis. This will be conducted on the FAS based on randomized treatment groups. The corresponding estimand for the supplementary analysis is the difference in the proportion of PsA patients who achieved ACR20 response at Week 12, regardless of whether the subject had discontinued study drug by Week 12, comparing each upadacitinib dose group vs placebo. The analysis will be conducted for those who are randomized, received study drug and have the efficacy measurement at Week 12 visit.

In addition to the supplementary analysis based on AO data, to explore various missing data assumptions including missing not at random (MNAR), tipping point analysis will also be performed for the primary endpoint by multiple imputations using logistic regression, allowing the imputed ACR response rate to systematically vary from 0% to 100% in both upadacitinib and placebo, respectively. Details of the tipping point analysis are outlined in Appendix B. Supportive analysis will also be conducted on the Per Protocol Analysis Set using the CMH model and NRI data handling.

9.2.3 Key Secondary Efficacy Analyses

The following is a list of ranked key secondary endpoints (upadacitinib versus placebo if not otherwise specified):

1. Change from baseline in HAQ-DI at Week 12;
2. Static Investigator Global Assessment (sIGA) of Psoriasis of 0 or 1 and at least a 2-point improvement from baseline at Week 16 (for subjects with baseline sIGA ≥ 2);

3. Psoriasis Area Severity Index (PASI) 75 response at Week 16 (for subjects with ≥ 3% BSA psoriasis at baseline);
4. Change from baseline in modified PsA Sharp/van der Heijde Score (SHS) at Week 24;
5. Proportion of subjects achieving MDA at Week 24;
6. Proportion of subjects with resolution of enthesitis (LEI = 0) at Week 24 (for subjects with baseline presence of enthesitis (LEI > 0));
7. ACR 20 response rate at Week 12 (non-inferiority of upadacitinib vs adalimumab);
8. Change from baseline in SF-36 PCS at Week 12;
9. Change from baseline in FACIT-Fatigue at Week 12;
10. ACR 20 response rate at Week 12 (superiority of upadacitinib vs. adalimumab);
11. Proportion of subjects with resolution of dactylitis (LDI = 0) at Week 24 (for subjects with baseline presence of dactylitis (LDI > 0));
12. Change from baseline in Patient's Assessment of Pain NRS at Week 12 (superiority of upadacitinib vs. adalimumab);
13. Change from baseline in HAQ-DI at Week 12 (superiority of upadacitinib vs. adalimumab);
14. Change from baseline in Self-Assessment of Psoriasis Symptoms (SAPS) at Week 16.

Additional key secondary efficacy endpoints are:

1. ACR50/70 response at Week 12;
2. ACR20 response at Week 2.

Binary Endpoints

For binary endpoints, frequencies and percentages will be reported for each treatment group. For the ranked secondary endpoints at the time point of interest, the primary estimand is the same as that for the primary efficacy endpoint as defined in Section 9.2.1, except for the definition of the efficacy measurement.

For the key secondary endpoints applicable only for sub-population, the estimands will be constructed similarly but based on the sub-population, as defined in the below table.

Endpoint	Sub-population
Static Investigator Global Assessment (sIGA) of Psoriasis of 0 or 1 and at least a 2 point improvement from baseline at Week 16	Subjects with baseline sIGA ≥ 2 are included in the analysis
Psoriasis Area Severity Index (PASI) 75 response at Week 16	Subjects with ≥ 3% BSA psoriasis at baseline are included in the analysis
Proportion of subjects with resolution of enthesitis (LEI = 0) at Week 24	Subjects with baseline LEI > 0 are included in the analysis
Proportion of subjects with resolution of dactylitis at Week 24	Subjects with baseline LDI > 0 are included in the analysis

For MDA at Week 24, resolution of enthesitis at Week 24 and resolution of dactylitis at Week 24, in addition to the NRI data handling as defined in Section 9.1.2, subjects who meet the rescue criteria (i.e., not achieving at least 20% improvement in either or both tender joint count (TJC) and swollen joint count (SJC) at both Week 12 and Week 16) will also be treated as non-responders. The corresponding primary estimand is the difference in the proportion of patients who achieved MDA/resolution of enthesitis/resolution of dactylitis at Week 24, and who did not discontinue study drug and did not initiate rescue therapy by Week 24. The comparison is for each upadacitinib dose group vs placebo for patients randomized and treated with at least one dose of study drug. NRI data handling will be used to analyze the primary estimand for binary endpoints.

Supplementary analysis using AO data handling will also be conducted. The corresponding supplementary estimand is the same as defined in Section 9.2.2 except for the definition of the efficacy measurement.

Point estimate and 95% CI using normal approximation will be provided for the response rate for each randomized treatment group. Treatment comparisons will be made between each upadacitinib dose and the combined placebo group using the Cochran-Mantel-Haenszel test (refer to Appendix A). The CMH test adjusts for the main stratification factor of current DMARD use (yes/no). Point estimate, 95% CI using normal approximation and the p-value for the treatment difference will be presented.

For ACR20 response at Week 12, analysis will be conducted to assess the non-inferiority of each upadacitinib dose versus adalimumab on the placebo-subtracted treatment difference using Koch's 3-arm test statistic as described in Appendix C. The test aims to show that upadacitinib preserves at least 50% of the placebo-subtracted adalimumab effect. If statistical significance is achieved for the non-inferiority assessment, superiority of upadacitinib versus adalimumab will be tested for ACR20 subsequently as described in Section 9.2.5.

Non-Radiographic Continuous Endpoints

For non-radiographic continuous endpoints, the estimand is the difference in mean change from baseline at the protocol defined time point (e.g., HAQ-DI at Week 12) under the assumption that patients with missing data, including those due to premature discontinuation of study drug, can have their measurements at the protocol defined time point predicted by their observed data and the observed data for other patients for their respective assessments during follow-up. The comparison is for each upadacitinib dose group vs placebo for patients randomized and treated with at least one dose of study drug.

To analyze the primary estimand for the continuous ranked secondary efficacy endpoints, statistical inference will be conducted using the MMRM model and the associated data handling as described in Section 9.1.2, with fixed effects of treatment, visit, treatment-by-

visit interaction, the stratification factor of current DMARD use (yes/no) and the continuous fixed covariate of baseline measurement. The LS mean and 95% CI for each randomized treatment group and the LS mean treatment difference, the associated 95% CI and the p-value between each upadacitinib dose group and the combined placebo group will be provided. For change from baseline in Patient's Assessment of Pain NRS and change from baseline in HAQ-DI at Week 12, superiority of upadacitinib vs. adalimumab will also be tested using the same MMRM model.

The supplementary analysis for ranked non-radiographic secondary continuous endpoints will be conducted using AO data handling and using the analysis of covariance (ANCOVA) model with treatment and the stratification factor of current DMARD use (yes/no) as the fixed factors and the corresponding baseline values as the covariates. The corresponding estimand is the difference in the mean change from baseline in the efficacy endpoints at the protocol defined time point regardless of treatment discontinuation or use of rescue medication, comparing each upadacitinib dose group vs placebo. The analysis will be conducted for those who are randomized, received study drug and have the efficacy measurement at the protocol defined time point.

For the key continuous endpoint change from baseline in HAQ-DI at Week 12, tipping point analysis will also be conducted using multiple imputation (MI) as additional supplementary analysis to explore various missing data assumptions including missing not at random (MNAR). Details of the tipping point analysis for continuous endpoints are outlined in Appendix D.

Radiographic Endpoints

For change from baseline in SHS at Week 24, statistical inference will be conducted using ANCOVA model and the associated data handling as described in Section 9.1.2 with treatment and the stratification factor of current DMARD use (yes/no) as the fixed factors and the corresponding baseline values as the covariates. The primary estimand for the x-ray assessment is the difference in the mean change from baseline in SHS score at Week 24, under the assumption that patients with missing data, including those due to

premature discontinuation of study drug or use of rescue medication, can have their measurements at Week 24 predicted using their SHS score from the baseline window and the Week 16 window. The comparison will be between each upadacitinib dose group vs placebo for patients randomized and treated with at least one dose of study drug.

To incorporate the uncertainty of a single imputation, the Multiple Imputation (MI) analysis will be conducted as a sensitivity analysis. The proposed MI model for the change from baseline in SHS score will be based on the linearity assumption for the subjects with missing data including those due to premature discontinuation of study drug or meeting the rescue criteria. More specifically, the regression method will be used where missing Week 24 observations will be imputed by regression upon all previous visits including baseline and Week 16, stratified by treatment group. The imputation model will also include baseline demographics and key baseline characteristics as appropriate. Analysis of covariance (ANCOVA) will be performed on each of the multiple imputed datasets. The results will be aggregated across the multiple imputed datasets using Rubin's method. The detailed SAS procedure are outlined in Appendix D.

The supplementary analysis will be performed using AO data handling with the same ANCOVA model. The corresponding estimand for the AO analysis is the difference in the mean change from baseline in SHS score at Week 24, regardless of premature discontinuation of study drug or use of rescue medication, comparing each upadacitinib dose group vs placebo. The analysis will be conducted for those who are randomized, received study drug and have the efficacy measurement at Week 24 visit.

The LS mean and 95% CI for each randomized treatment group and the LS mean treatment difference, the associated 95% CI and the p-value between each upadacitinib dose group and the combined placebo group will be provided.

In another sensitivity analysis, a random coefficient model will be applied, in which SHS measurements are used as the response variable, including measurements at Baseline and post-Baseline. The model is specified with a random intercept and a random slope on time, and includes treatment, current DMARD use (yes/no), and treatment-by-time

interaction in the model. The "group" option in PROC MIXED is used to allow for heterogeneity among treatment groups in the covariance structure. The coefficient for treatment-by-time interaction is used to estimate the difference in rate of progression (per 24 weeks) comparing upadacitinib and placebo. The rate of progression (per 24 weeks) for each group respectively is also estimated using the coefficients for time and treatment-by-time interaction. In case the random coefficient model does not converge, a simplified model may be explored by removing the random intercept or fitting a fixed-effect linear effect model.

Additionally, in the event that data severely deviates from the normal distribution, non-parametric analyses such as the ranked ANCOVA may be considered for treatment comparison.

9.2.4 Efficacy Analyses for Additional Endpoints

Additional efficacy analysis includes the following endpoints at the scheduled time points in Section 5.2 other than those specified for the primary and key secondary variables:

- Proportion of subjects with no radiographic progression is defined as
  - change from baseline in SHS ≤ 0;
  - change from baseline in SHS ≤ 0.5.
- Change from baseline in joint space narrowing score and joint erosion score.
- Change from baseline in individual components of ACR response
  - Change from baseline in Tender Joint Count (TJC) (0 – 68);
  - Change from baseline in Swollen Joint Count (SJC) (0 – 66);
  - Change from baseline in Physician Global Assessment (PGA) – Disease Activity (NRS);
  - Change from baseline in Patient's Global Assessment (PtGA) – Disease Activity (NRS);
  - Change from baseline in Patient's Assessment of Pain Numerical Rating Scale (NRS);

  - Change from baseline in Health Assessment Questionnaire – Disability Index (HAQ-DI);
  - Change from baseline in High-Sensitivity C Reactive Protein (hs-CRP);
- ACR 20/50/70 response rates;
- Proportion of subjects achieving MDA;
- Change from baseline in Leeds Dactylitis Index (LDI) (for subjects with baseline LDI > 0);
- Change from baseline in dactylitis count (for subjects with baseline LDI > 0);
- Proportion of subjects with resolution of dactylitis (for subjects with baseline LDI > 0);
- Change from baseline in LEI (for subjects with baseline LEI > 0);
- Proportion of subjects with resolution of enthesitis defined as LEI = 0 (for subjects with baseline LEI > 0);
- Change from baseline in SPARCC Enthesitis Index (for subjects with baseline SPARCC Enthesitis Index > 0);
- Proportion of subjects with resolution of enthesitis defined as SPARCC Enthesitis Index = 0 (for subjects with baseline SPARCC Enthesitis Index > 0);
- Change from baseline in total enthesitis count (for subjects with baseline total enthesitis count > 0);
- Proportion of subjects with resolution of enthesitis defined as total enthesitis count = 0 (for subjects with baseline total enthesitis count > 0);
- PASI 75/90/100 response rates (for subjects with ≥ 3% Body Surface Area (BSA) psoriasis at baseline);
- Proportion of subjects achieving a static Investigator Global Assessment of psoriasis (sIGA) score of 0 or 1 and at least a 2-point improvement from baseline (for subjects with baseline sIGA ≥ 2);
- Change from baseline in BSA-Ps (for subjects with baseline BSA-Ps > 0);
- Modified Psoriatic Arthritis Response Criteria (PsARC) response rate;
- Change from baseline in Disease Activity Score 28 (DAS28) (CRP);
- Change from baseline in DAS28 (ESR);

- Change from baseline in PsA Disease Activity Score (PASDAS);
- Change from baseline in Disease Activity In Psoriatic Arthritis (DAPSA) score;
- Change from baseline in Short Form 36 (SF-36) Health;
- Change from baseline in Functional Assessment of Chronic Illness Therapy (FACIT)-Fatigue;
- Change from baseline in EuroQol-5D-5L (EQ-5D-5L) index and VAS score;
- Change from baseline in Work Productivity and Activity Impairment (WPAI);
- Cumulative Health Resource Utilization (HRU);
- Change from baseline in Self-Assessment of Psoriasis Symptoms (SAPS);
- Change from baseline in Bath Ankylosing Spondylitis Disease Activity Index (BASDAI) (out of subjects with presence of psoriatic spondylitis at baseline);
- BASDAI 50 response rates (out of subjects with presence of psoriatic spondylitis at baseline);
- Change from baseline in Morning stiffness (mean of BASDAI Questions 5 and 6);
- Change from baseline in Morning stiffness severity 0 - 10 (BASDAI Question 5)
- Change from baseline in Morning stiffness duration 0 - 10 (BASDAI Question 6)
- Change from baseline in Ankylosing Spondylitis Disease Activity Score (ASDAS) (out of subjects with presence of psoriatic spondylitis at baseline);
- Proportion of subjects with ASDAS Inactive Disease (out of subjects with presence of psoriatic spondylitis at baseline);
- Proportion of subjects with ASDAS Major Improvement (out of subjects with presence of psoriatic spondylitis at baseline);
- Proportion of subjects with ASDAS Clinically Important Improvement (out of subjects with presence of psoriatic spondylitis at baseline);
- Proportion of subjects achieving a clinically meaningful improvement in HAQ-DI (≥ 0.35).

Non Radiographic Endpoints

For binary endpoints, frequencies and percentages will be reported for each treatment group. NRI will be used for primary analysis. For enthesitis and dactylitis related endpoints at Week 24, in addition to the NRI data handling as defined in Section 9.1.2, subjects who meet the rescue criteria (i.e., not achieving at least 20% improvement in either or both tender joint count (TJC) and swollen joint count (SJC) at both Week 12 and Week 16) will also be treated as non-responders. AO will be used as supplementary analysis. Point estimate and 95% CI using normal approximation will be provided for the response rate for each randomized treatment group. Treatment comparisons will be made between each upadacitinib dose to the combined placebo group and to the adalimumab group using the CMH (refer to Appendix A) test adjusting for the main stratification factor of current DMARD use (yes/no). Point estimate, 95% CI using normal approximation and the nominal p-value for the treatment difference will be presented.

For non-radiographic continuous endpoints, the LS mean and 95% CI for each randomized treatment group and the LS mean treatment difference, the associated 95% CI and the p-value between each upadacitinib dose group to the combined placebo group and to the adalimumab group will be provided using MMRM model as described in Section 9.1.2, with fixed effects of treatment, visit, treatment-by-visit interaction, the stratification factor of current DMARD use (yes/no) and the continuous fixed covariates of baseline measurement. The nominal p-value will be provided. For enthesitis and dactylitis related endpoints at Week 24, exploratory analysis accounting for potential influential outliers may be performed.

For change from baseline in each of the seven ACR components, the ANCOVA model using AO data handling as described in Section 9.2.3 will be conducted as supplementary analysis, with treatment and the stratification factor of current DMARD use (yes/no) as the fixed factors and the corresponding baseline values as the covariates. Nominal p-values will be provided.

Radiographic Endpoints

For proportion of subjects with no radiographic progression at Week 24, point estimate and 95% CI of the response rate for each randomized treatment group will be provided. Comparisons will be made between the upadacitinib 15 mg QD and 30 mg QD group and the combined placebo group using the CMH (refer to Appendix A) test adjusting for current DMARD use (yes/no). Point estimate, 95% CI and nominal p-value for the treatment comparison will be presented. Non-progression will be derived using both linear extrapolation imputation and AO.

For change from baseline in joint space narrowing score and joint erosion score at Week 24, linear extrapolation analysis and AO analysis will be performed similarly as described in Section 9.2.3.

In the linear extrapolation analysis for Week 56 radiographic endpoints, the missing Week 56 data will be imputed via linear extrapolation using available post-baseline x-ray data in the latest analysis window prior to Week 56. For placebo subjects who switched to upadacitinib at Week 24, their Week 56 x-ray data will be imputed via linear extrapolation using Week 24 x-ray data. In the AO analysis, the observed Week 56 measurements will be used regardless of treatment switching. Similar statistical inference approach as Week 24 analysis by original randomized treatment group will be used.

Nominal p-values will be provided.

9.2.5 Handling of Multiplicity

The overall type I error rate of the primary and ranked key secondary endpoints for the two doses will be strongly controlled using a graphical multiple testing procedure [1]. Specifically, the testing will utilize the endpoint sequence of primary endpoint followed by the ranked key secondary endpoints in the order as specified in Section 9.2.3, and will begin with testing the primary endpoint using $\alpha/2$ for each dose ($\alpha = 0.0499$ accounting for 0.0001 spending after interim futility analysis). Continued testing will follow a pre-

specified α transfer path which includes downstream transfer along the endpoint sequence within each dose as well as cross-dose transfer.

The graph for the testing procedures is provided in Figure 2. In the graph, the arrows specify the α transfer paths. Once an endpoint is rejected (i.e., deemed significant) at its assigned significance level, its significance level will be transferred to subsequent endpoint(s) following the arrow(s). The numbers on the arrows denote the weights for transferring significance levels. Specifically, the weight 1 denotes 100% transfer of significance level.

Figure 2. Graphical Multiple Testing Procedure

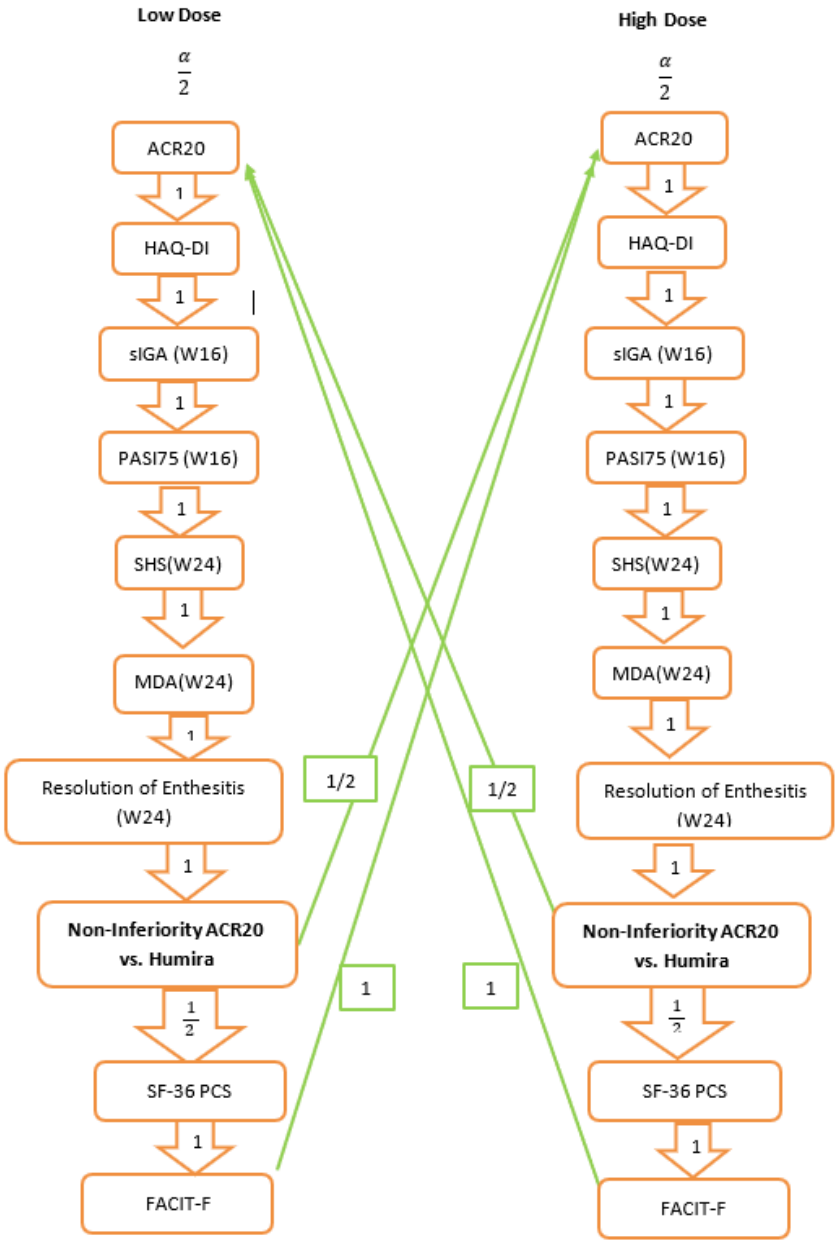
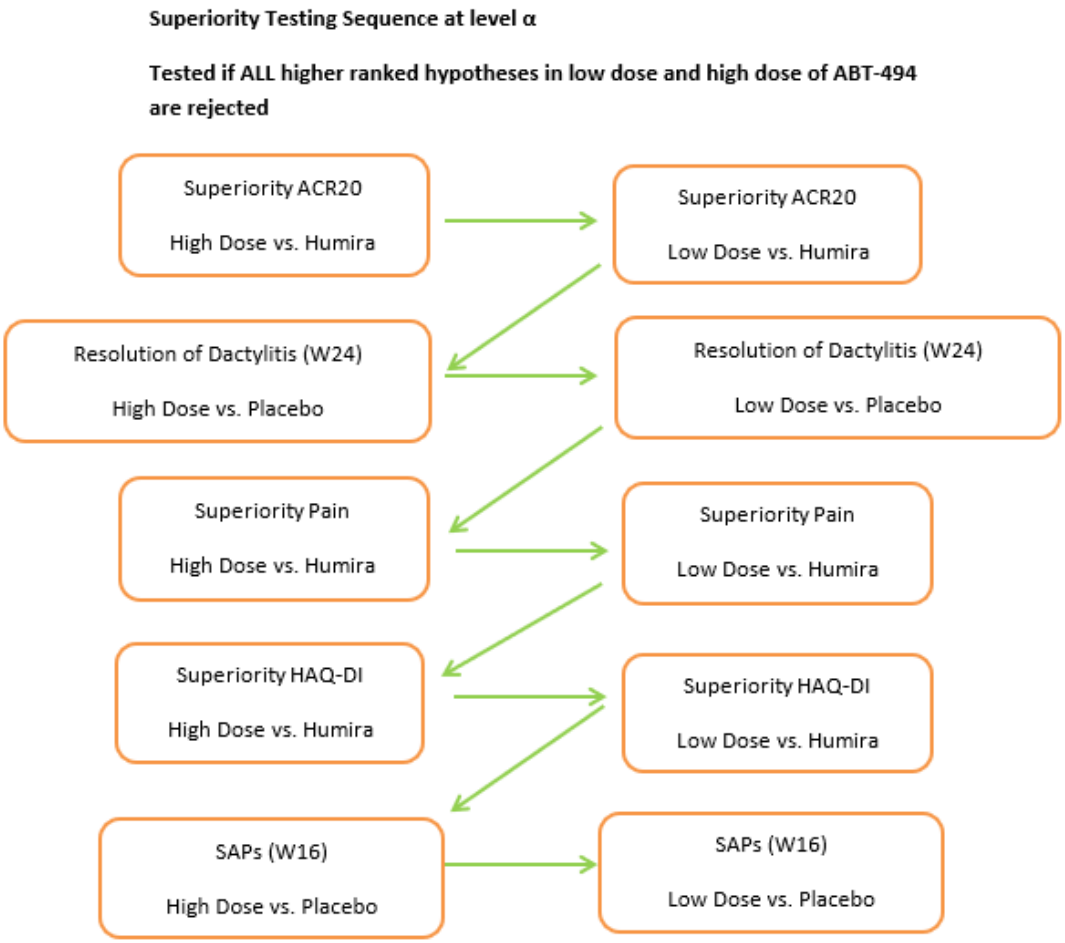


Figure 2. Graphical Multiple Testing Procedure (Continued)



9.2.6 Efficacy Subgroup Analysis

The primary efficacy endpoint will be examined in the subgroups listed in Table 7 below. Treatment difference between each upadacitinib dose and the combined placebo group will be presented with point estimate and 95% confidence interval using normal approximation. No p-value will be provided for subgroup analysis.

Table 7. Subgroups for Efficacy Analysis

Subgroup Factor	Categories
Age	< 65, ≥ 65 - < 75, ≥ 75 years (if sample size allows)
Sex	Male or Female
BMI	< 25 kg/m^2 or ≥ 25 kg/m^2
Race	White, Non-white
Geographic region	North America, Western Europe and Oceania, Eastern Europe, Latin-America, Asia, and Other
Duration of PsA diagnosis	≤ 5, > 5 - ≤ 10, > 10 years
Baseline hsCRP	≤ ULN or > ULN
Number of prior non-biologic DMARDs	≤ 1, > 1
Current use of non-biologic DMARDs	Yes or No

9.3 Long-Term Efficacy Analysis

Assessments to evaluate long-term efficacy will be analyzed for the following measures at Week 2, 4, 8, 12, 16, 20, 24, 28, 32, 36, 44, 56 and every 12 weeks thereafter until completion of the study:

- ACR20/50/70 response rates
- Change from baseline in individual ACR components
- Change from baseline in DAS28 (CRP) and DAS28 (ESR)
- PsARC response rate
- Change from baseline in DAPSA
- Proportion of subjects achieving a clinically meaningful improvement in HAQ-DI (≥ 0.35)

Assessments to evaluate long-term efficacy will be analyzed for the following measures at Week 12, 24, 36, 56 and every 12 weeks thereafter until completion of the study:

- Change from baseline in EQ-5D-5L index and VAS score
- Change from baseline in FACIT-Fatigue
- Change from baseline in SF-36

- Change from baseline in WPAI
- Change from baseline in BASDAI
- Cumulative HRU
- Change from baseline in PASDAS
- BASDAI 50 response rates
- Change from baseline in ASDAS
- Proportion of subjects with ASDAS Inactive Disease
- Proportion of subjects with ASDAS Major Improvement
- Proportion of subjects with ASDAS Clinically Important Improvement

Assessments to evaluate long-term efficacy will be analyzed for the following measures at Week 12, 16, 24, 36, 56 and every 12 weeks thereafter until completion of the study:

- Change from baseline in BSA-Ps
- PASI 75/90/100 response rates
- Static Investigator Global Assessment (sIGA) of Psoriasis of 0 or 1 and at least a 2-point improvement from baseline
- Proportion of subjects achieving Minimal Disease Activity (MDA)
- Change from baseline in LDI
- Change from baseline in LEI
- Change from baseline in SPARCC enthesitis index
- Change from baseline in dactylitis count
- Proportion of subjects with resolution of dactylitis
- Proportion of subjects with resolution of enthesitis sites included in the LEI
- Proportion of subjects with resolution of enthesitis sites included in the SPARCC Enthesitis Index
- Change from baseline in total enthesitis count
- Proportion of subjects with resolution of enthesitis

Assessments to evaluate long-term efficacy will be analyzed for the following measures at Week 16, 24, 36, 56 and every 12 weeks thereafter until completion of the study:

- Change from baseline in SAPS

Additionally, assessments to evaluate long-term radiographic change will be analyzed for the following measures at Week 24, 56, 104, 152 and 260/PD visits:

- Change from baseline in SHS
- Change from baseline in joint space narrowing score and joint erosion score
- Proportion of subjects with no radiographic progression (defined as change from baseline in SHS ≤ 0);

Descriptive statistics based on AO will be provided for each randomized treatment group sequence as defined in Section 9.1.1. These include the number of observations, mean, standard deviation, 95% CI, median, minimum and maximum for continuous endpoints; and frequencies and percentages with 95% CI using normal approximation for binary endpoints. Plot for each randomized treatment group sequence over time will be provided for key endpoints. Additionally, treatment comparisons between each upadacitinib dose versus adalimumab will be made for efficacy endpoints for the originally randomized upadacitinib groups and adalimumab group. Point estimate, nominal p value and 95% CI will be provided up to Week 56. Treatment comparisons may also be provided after Week 56 as appropriate.

In addition, NRI approach and/or GLMM will be used for non-radiographic binary endpoint analysis, and MMRM will be used for continuous endpoint analysis. The analyses will be performed for the treatment sequences. The longitudinal analysis using MMRM or GLMM was described in Section 9.1.2. Point estimates and 95% CI from the model will be provided for each treatment sequence. Plot for each randomized treatment group sequence over time will be provided for key endpoints.

For radiographic related endpoints, the analysis will be based on linear extrapolation imputation and AO data. MMRM described in Section 9.1.2 will be conducted as mixed model based on AO data for Week 104 and future analyses.

9.4 Efficacy Variables Definitions and Conventions

9.4.1 ACR Criteria

ACR criteria are a commonly used standard criteria set mentioned in the FDA industry guidance to evaluate the effectiveness of investigational drug in reduction of disease activity. It is a composite measurement calculated based on the improvement over a set of core measurements.

ACR20 is defined as at least 20% improvement (compared to baseline values) in tender and swollen joint counts and at least 20% improvement in 3 of the remaining 5 core set measures (subject assessment of pain, subject global assessment of disease activity, physician global assessment of disease activity, subject assessment of physical function and acute phase reactant hsCRP).

ACR50 and ACR70 are similarly defined with at least 50% and 70% improvement, respectively.

A subject will be classified as an ACR20 (ACR50, ACR70) responder, if the following conditions are met:

1. ≥ 20% (50%, 70%) improvement from baseline in tender joint count (TJC68) and
2. ≥ 20% (50%, 70%) improvement from baseline in swollen joint count (SJC66) and
3. ≥ 20% (50%, 70%) improvement from baseline in at least 3 of the following 5:
   - patient's assessment of pain
   - patient's global assessment of disease activity (PtGA)
   - physician's global assessment of disease activity (PGA)

- patient's self-assessment of physical function (i.e., measured by Health Assessment Questionnaire HAQ-DI score)
- Acute-phase reactant value CRP

Seven components are included in the ACR response criteria. Missing values for a component can occur due to data collection issues for a particular form, a missed visit or dropout from the study. Depending on the pattern of the missing components, ACR responses may be or may not be determined for a visit date using (partially) observed values only. In the case when ACR responses cannot be determined for any visit date within a visit window, partially observed data from different visit dates within the same visit window can be combined to determine ACR responses for the visit window.

To maximize the utilization of observed information at certain visits and be scientifically as robust as possible, the principle to calculate ACR response is to minimize imputation whenever possible. "As Observed" ACR response will be calculated first based on a derived visit window instead of the nominal visit identifier (e.g., Week 6 visit) collected from the CRF.

To calculate "as observed" ACR responses:

Identify the observed component 20% improvement indicator (0/1/missing), 1 means achieving ≥ 20% improvement from baseline and 0 means < 20% improvement from baseline.

ACR20 = 0 if TJC indicator = 0 OR SJC indicator = 0 OR at least 3 out of 5 components improvement indicators = 0;

ACR20 = 1 if TJC indicator = 1 AND SJC indicator = 1 AND at least 3 out of 5 components improvement indicators = 1.

For all other cases, "as observed" ACR20 = missing since ACR20 cannot be determined.

The following table illustrates examples for as-observed ACR calculations.

Example	TJC 68	SJC 66	Component 1	Component 2	Component 3	Component 4	Component 5	ACR20-Response?
A	1	1	1	1	1	.	.	Yes
B	1	0	1	1	1	1	1	No
C	.	0	.	.	.	.	.	No
D	1	.	1	1	1	1	1	.
E	1	1	0	0	0	1	1	No
F	.	.	0	0	0	.	.	No
G	1	1	1	1	0	0	.	.

Legend: 1 = ≥ 20% improved compared to baseline; 0 = < 20% improved compared to baseline; "." missing

Derived visit windowing Rule for ACR Response Calculation:

To identify the component value in a visit window:

ACR component values will first be determined at each date within a visit window.

ACR component values at each date will be combined to determine the "as observed" ACR composite score at each date in each window.

After this calculation, if multiple non-missing ACR composite scores are available within a given visit window, the non-missing ACR composite score closest to the target day will be used. If two composite scores have the same distance from the target day, the later one will be used. The corresponding date will be used as the "observed" ACR response date in the derived efficacy dataset.

If a non-missing ACR composite score is not available for any day within a given visit window, the windowed component values for that visit will be used to calculate the ACR composite score for that visit window (component value windowing follow the same rules as in steps described above). Date of TJC will be used as the "observed" ACR date in the derived efficacy dataset.

The following rules are applied for the NRI imputation for the ACR responses.

Non-Responder Imputation (NRI):

- **Step 1:** all missing components will be imputed using LOCF, and then the ACR composite score can be calculated
- **Step 2:** if the ACR composite score cannot be determined by step 1, the ACR composite score will be imputed as 0. In addition, subjects who prematurely discontinue from the study drug will be considered as non-responders (ACR = 0) for all subsequent visits after the discontinuation date.

9.4.2 Joint Evaluation

Anatomical joints are evaluated for swelling and tenderness at every study visit. The 34 anatomical joints in Table 8 are assessed in this study for both the left and right side of the body.

Table 8. Anatomical Joints Assessed for Calculation of Tender and Swollen Joint Counts (TJC68 and SJC66)

Temporomandibular	Sternoclavicular	Acromio-clavicular	Shoulder
Elbow	Wrist	Metacarpophalangeal I	Metacarpophalangeal II
Metacarpophalangeal III	Metacarpophalangeal IV	Metacarpophalangeal V	Thumb Interphalangeal
Proximal Interphalangeal II	Proximal Interphalangeal III	Proximal Interphalangeal IV	Proximal Interphalangeal V
Distal Interphalangeal II	Distal Interphalangeal III	Distal Interphalangeal IV	Distal Interphalangeal V
Hip[a]	Knee	Ankle	Tarsus
Metatarsophalangeal I	Metatarsophalangeal II	Metatarsophalangeal III	Metatarsophalangeal IV
Metatarsophalangeal V	Great Toe/Hallux	Interphalangeal II	Interphalangeal III
Interphalangeal IV	Interphalangeal V		

a. Hip joints are not assessed for swelling.

At each study visit, a joint evaluator assessed whether a particular joint was "tender or painful" where presence of tenderness was scored as "1" and the absence of tenderness was scored as "0," provided the joint was not replaced ("9") or could not be assessed ("NA") due to other reasons (e.g., post-corticosteroid joint injection). The total tender joint count (TJC68), which is based on 68 joints, will be derived as the sum of all "1s" and proportional extrapolation will be used to impute joint counts for the joints that are

replaced or not assessed. A similar method will be followed for the derivation of total swollen joint count (SJC66), which is based on 66 joints as the hip joints are excluded. Thus, the range for TJC68 will be 0 to 68 and 0 to 66 for SJC66.

9.4.3 Patient's Global Assessment of Disease Activity Numeric Rating Scale (NRS)

The subject will assess his/her disease activity using a Patient's Global Assessment of Disease NRS. The range is 0 to 10 with no activity being indicated by 0 and severe activity by 10.

9.4.4 Physician's Global Assessment of Disease Activity Numeric Rating Scale (NRS)

The physician will assess Patient's disease activity at the time of visit using a Physician's Global Assessment of Disease NRS. The range is 0 to 10 with no activity being indicated by 0 and severe activity by 10.

9.4.5 Patient's Assessment of Pain Numeric Rating Sale (NRS)

The subject will assess his/her pain using the Patient's Assessment Pain NRS. The range is 0 to 10 with no activity being indicated by 0 and severe activity by 10.

9.4.6 Disease Activity Score (DAS28)

DAS28 (CRP) and DAS28 (ESR) are composite indices to assess disease activity in RA using hsCRP or ESR measurement, respectively. The DAS provides a score between 0 and 10, indicating how active the rheumatoid arthritis is at the time of measurement.

DAS28 (CRP) and DAS28 (ESR) can be calculated based on Tender Joint Count, Swollen Joint Count, Patient's Global Assessment of Disease Activity (PtGA) (in mm), and hsCRP (in mg/L) or ESR (mm/hr).

$$DAS28\ (CRP) = 0.56 \times \sqrt{(TJC28^{*})} + 0.28 \times \sqrt{(SJC28^{**})} + 0.36 \times \ln(hsCRP^{\&} + 1) + 0.014 \times PtGA^{\gg} + 0.96$$

$$DAS28\ (ESR) = 0.56 \times \sqrt{(TJC28^{*})} + 0.28 \times \sqrt{(SJC28^{**})} + 0.70 \times \ln(ESR^{\#}) + 0.014 \times PtGA^{\gg}$$

where $\sqrt{}$ is square root and ln is natural log.

* TJC28 refers to the Subject's total Tender Joint Count out of the provided 28 evaluated joints.

** SJC28 refers to the Subject's total Swollen Joint Count out of the provided 28 evaluated joints.

& hsCRP refers to the high-sensitivity c-reactive protein lab value. hsCRP unit in the DAS28 (CRP) equation is expressed as mg/L.

\# ESR refers to the Erythrocyte sedimentation rate. ESR unit in the DAS28 (ESR) equation is expressed as mm/hr.

» PtGA refers to the Patient's Global Assessment of Disease Activity.

Table 9. Anatomical Joints for DAS28 (CRP) Calculation

Shoulder	Elbow	Wrist	Thumb Interphalangeal
Metacarpophalangeal I	Metacarpophalangeal II	Metacarpophalangeal III	Metacarpophalangeal IV
Metacarpophalangeal V	Proximal Interphalangeal II	Proximal Interphalangeal III	Proximal Interphalangeal IV
Proximal Interphalangeal V	Knee		

As Patient's Global Assessment of Disease Activity is collected with the scale of 0 - 10 NRS, the variable needs to be multiplied by 10 before being used the DAS28 formula.

To calculate observed DAS28 scores, the observed component value will be calculated first. Then the components will be included in the calculation per the DAS formula selected. If any observed component is missing in a window, then the observed DAS28 score will be missing.

9.4.7 Disease Activity in Psoriatic Arthritis (DAPSA) Score

DAPSA is a continuous endpoint that measures the disease activity in psoriatic arthritis. DAPSA consists of five components: Tender Joint Count 68, Swollen Joint Count 66,

Patient's Assessment of Pain (0 - 10 NRS), Patient's Global Assessment of Disease Activity (0 - 10 NRS), and hsCRP (in mg/dL).

DAPSA = SJC66 + TJC68 + Pt pain (0 - 10 NRS) + PtGA (0 - 10 NRS) + CRP (in mg/dL)

To calculate observed DAPSA scores, the observed component value will be calculated first. Then the components will be included in the calculation per the DAPSA formula. If any observed component is missing in a window, then the observed DAPSA score will be missing.

9.4.8 PsA Disease Activity Score (PASDAS)

PASDAS is a continuous scale of combined joint, dactylitis and enthesitis assessments, physician and patient global assessments for arthritis, SF36-PCS, and hsCRP measurements.

PASDAS = (((0.18 √(PGA)) + 0.159 √(PtGA) – 0.253 √(SF36-PCS) + 0.101 ln (SJC66 + 1) + 0.048 ln (TJC68 + 1) + 0.23 ln (Leeds Enthesitis Index + 1) + 0.37 ln (Tender Dactylitis Count + 1) + 0.102 ln (hsCRP + 1) + 2) * 1.5,

where √ is square root and ln is natural log. PtGA is on the scale of 0 – 100 and PGA is on the scale of 0 – 100. As PtGA and PGA are collected with the scale of 0 - 10 NRS, their values need to be multiplied by 10 before being used in the PASDAS formula.

SF36-PCS is the physical component scale in SF36 instrument. The unit for hsCRP is mg/L.

The tender dactylitis count is based on the dactylitis assessment. The count can be calculated by summing the digits with a "presence of tenderness" score = 1.

The Leeds Enthesitis Index (LEI) evaluates enthesitis at 6 most commonly involved entheseal sites, as indicated in the table below. The LEI is calculated by taking the sum of the scores from the 6 sites. The LEI ranges from 0 to 6.

		Tenderness in Left			Tenderness in Right		
		YES = 1	NO = 0	Not assessed = NA	YES = 1	NO = 0	Not assessed = NA
1	Lateral epicondyle						
2	Achilles tendon insertion						
3	Medial femoral condyle						

To calculate observed PASDAS scores, the observed component value will be calculated first. Then the components will be included in the calculation per the PASDAS formula. If any observed component is missing in a window, then the observed PASDAS score will be missing.

9.4.9 Disability Index of Health Assessment Questionnaire (HAQ DI)

HAQ-DI is a self-reported patient outcome measurement. It is calculated as the mean of the scores from 8 following categories with a range 0 – 3: Dressing and Grooming, Rising, Eating, Walking, Hygiene, Reach, Grip, and Activities. Higher scores reflect greater disability.

The maximum score for all the questions in each category is considered as the score for the category. The HAQ-DI takes into account the subject's use of aids or devices or assistance in the scoring algorithm for a disability category. For each of the eight disability categories there is an AIDS OR DEVICES companion variable(s) that is used to record the type of assistance, if any, a subject uses for his/her usual activities. If aids or devices and/or assistance from another person are checked for a disability category, the score for this category is set to 2 (much difficulty), if the original score is 0 (no difficulty) or 1 (some difficulty). The HAQ-DI is then calculated by summing the adjusted categories scores and dividing by the number of categories answered. The HAQ-DI cannot be calculated if the patient does not have scores for at least 6 categories.

9.4.10 Static Investigator Global Assessment of Psoriasis (sIGA)

The sIGA is a 5 point score ranging from 0 to 4, based on the investigator's assessment of the average elevation, erythema, and scaling of all psoriatic lesions. The assessment is considered "static" which refers to the patients disease state at the time of the assessments, without comparison to any of the patient's previous disease states, whether at Baseline or at a previous visit. A lower score indicates less severe psoriasis (0 = clear, 1 = almost clear, 2 = mild, 3 = moderate and 4 = severe).

A binary clinical endpoint based on sIGA is considered in this study. It is the proportion of subjects achieving a sIGA score of 0 or 1 and at least a 2-point improvement from baseline. This endpoint is calculated among the subjects with baseline sIGA score ≥ 2.

9.4.11 Psoriasis Area Severity Index (PASI)

Psoriasis Area Severity Index (PASI) has four anatomic sites – head, upper extremities, trunk, and lower extremities – which are assessed for erythema, induration and desquamation using a 5-point scale:

- 0 = no symptoms
- 1 = slight
- 2 = moderate
- 3 = marked
- 4 = very marked

Based on the extent of lesions in a given anatomic site, the area affected is assigned a numerical value:

- 0 = no involvement
- 1 = < 10%
- 2 = 10% – 29%
- 3 = 30% – 49%
- 4 = 50% – 69%

- 5 = 70% – 89%
- 6 = 90% – 100%

Since the head, upper extremities, trunk and lower extremities correspond to approximately 10, 20, 30 and 40% of body surface area, respectively; the PASI score is calculated using the formula:

$$PASI = 0.1(E_h + I_h + D_h)A_h + 0.2(E_u + I_u + D_u)A_u + 0.3(E_t + I_t + D_t)A_t + 0.4(E_l + I_l + D_l)A_l$$

where *E*, *I*, *D*, and *A* denote erythema, induration, desquamation, and area, respectively, and *h*, *u*, *t*, and *l* denote head, upper extremities, trunk, and lower extremities, respectively. PASI scores range from 0.0 to 72.0 with the highest score representing complete erythroderma of the severest possible degree. Typically scores of 3 or less represent mild disease, scores over 3 and up and including 15 represent moderate disease and scores over 15 are considered to be associated with severe disease.

If one item is missing, PASI is not scored.

PASI 75 (PASI 50, PASI 90, PASI 100) responder is reached if there is at least a 75% (50%, 90%, 100%) reduction in PASI score (≥ PASI 75/50/90/100 response) at a visit relative to the Baseline PASI score.

PASI is summarized in subjects with ≥ 3% BSA (Body Surface Area) psoriasis involvement at baseline.

9.4.12 Minimum Disease Activity for PsA

A patient is classified as in MDA when 5 of the following 7 criteria are met:

- TJC68 ≤ 1
- SJC66 ≤ 1
- PASI ≤ 1 or BSA-Ps ≤ 3%
- Patient's assessment of pain ≤ 1.5 (0 - 10 NRS)

- Patient's Global Assessment of disease activity ≤ 2 (0 - 10 NRS)
- HAQ-DI score ≤ 0.5
- Leeds Enthesitis Index ≤ 1

MDA response can be determined if at least 5 of the 7 criteria are met (responder), or if at least 3 of the 7 criteria are not met (non-responder). Selection of multiple MDA responses within one visit window follows the same rules as ACR. Missing values for each component can occur due to data collection issues for a particular form, a missed visit or dropout from the study. Depending on the pattern of the missing components, MDA responses may be or may not be determined for a visit date using (partially) observed values only. In the case when MDA responses cannot be determined for any visit date within a visit window, partially observed data from different visit dates within the visit window can be combined to determine MDA responses for the visit window.

9.4.13 Self-Assessment of Psoriasis Symptoms (SAPS)

The Self-Assessment of Psoriasis Symptoms (SAPS) contains 11 symptom-focused items. Each item is scored from 0 to 10, with 0 being least severe and 10 being most severe. The total score is generated by summing the 11 items. The total score ranges from 0 to 110.

9.4.14 Leeds Dactylitis Index (LDI) and Dactylitis Count

The Leeds Dactylitis Index (LDI) is a score based on finger circumference and tenderness, assessed and summed across all dactylitic digits. The presence of dactylitic digit is defined as at least one affected AND tender digit with circumference increase over reference digit ≥ 10%. For each of 20 digits of a subject, a digit final score needs to be calculated first. For an unaffected digit, the digit final score is set to be 0. For an affected digit, the digit final score is calculated as (A/B-1)*100*C if A/B ≥ 1.1, and digit final score = 0 if A/B ≤ 1.1, where A denotes the circumference of the digit, B the reference circumference, and C the tenderness score. The reference circumference can be either the circumference of the unaffected contralateral digit if available, or from a reference table if otherwise. LDI is the sum of the digit final scores over all 20 digits.

The dactylitis count will be calculated as the number of digits (hands and feet) with presence of dactylitis. The count ranges from 0 to 20.

The proportion of subjects with resolution of dactylitis is defined as the proportion of subjects with LDI = 0.

9.4.15 Leeds and SPARCC Enthesitis Indices and Total Enthesitis Count

For the Spondyloarthritis Research Consortium of Canada (SPARCC) enthesitis index 16 sites are evaluated as indicated in rows 1 – 8 in the table below. Tenderness on examination is recorded as either present (coded as 1), absent (coded as 0), or not assessed for each site. The SPARCC enthesitis index is calculated by taking the sum of the scores from the 16 sites. The SPARCC score ranges from 0 to 16.

The Leeds Enthesitis Index evaluates enthesitis at the 6 entheseal sites indicated in rows 2, 7 and 9 in the table below. Tenderness on examination is recorded as either present (coded as 1), absent (coded as 0), or not assessed for each of the 6 sites. The LEI is calculated by taking the sum of the scores from the 6 sites. The LEI ranges from 0 to 6.

The total enthesitis count is calculated by taking the sum of the tenderness scores from all 18 sites in the table below.

The proportion of subjects with resolution of enthesitis sites included in the LEI is defined as the proportion of subjects with LEI = 0; the proportion with resolution of the SPARCC Enthesitis Index and of the total enthesitis count are similarly defined.

		Tenderness in Left			Tenderness in Right		
		Present = 1	**Absent = 0**	**Not assessed = NA**	**Present = 1**	**Absent = 0**	**Not assessed = NA**
1	Medial epicondyle						
2	Lateral epicondyle						
3	Supraspinatus insertion into the greater tuberosity of humerus						
4	Greater trochanter						
5	Quadriceps insertion into superior border of patella						
6	Patellar ligament insertion into inferior pole of patella or tibial tubercle						
7	Achilles tendon insertion into calcaneum						
8	Plantar fascia insertion into calcaneum						
9	Medial femoral condyle						

9.4.16 Body Surface Area (BSA) – Psoriasis

The subject's right or left hand should be selected as the measuring device. For purposes of clinical estimation, the total surface of the palm plus five digits will be assumed to be approximately equivalent to 1%. Measurement of the total area of involvement by the physician is aided by imagining if scattered plaques were moved so that they were next to each other and then estimating the total area involved.

9.4.17 Bath Ankylosing Spondylitis Disease Activity Index (BASDAI) and Morning Stiffness Score

The BASDAI is composed of 6 items investigating 5 domains (fatigue, spinal pain, joint pain/swelling, areas of localized tenderness, morning stiffness), with 1 item for each of the first four domains and 2 items for the last domain (morning stiffness). Each item is scored on a 0 - 10 NRS. A lower score indicates less disease activity.

Scoring of the BASDAI is as follows:

1. Measure each item of the BASDAI in NRS (out of a total of 10)
2. BASDAI Score = 0.2*(Item1 + Item2 + Item3 + Item4 + 0.5*Item5 + 0.5* Item6)

The BASDAI Score ranges from 0 – 10. If one of the 5 items (Questions 1 – Question 4, inflammation) is missing, then the score is the mean of the 4 non-missing items (total of 4 non-missing items divided by 4). If more than 1 of the 5 items is missing, then the BASDAI score is missing.

Note: Question 5 and Question 6 jointly constitute Item 5 (inflammation). If both Questions 5 and 6 are missing, and questions 1 through 4 are non-missing, then only one item will be considered missing. The BASDAI score can still be calculated as the mean of Questions 1 – 4. However, if, for example, both Question 6 and Question 1 are missing, then 2 items will be considered missing, as the inflammation calculation would be incomplete. The BASDAI score would then be considered missing in this case.

The Morning Stiffness Score is the average of BASDAI questions 5 and 6 and it ranges from 0 – 10.

9.4.18 Modified Psoriatic Arthritis Response Criteria (PsARC)

The PsARC was developed as a PsA-specific composite responder index. To achieve response, a patient must achieve 2 of the following 4 items, one of which has to be a Tender Joint Count 68 or Swollen Joint Count 66, and no worsening of any measure:

- ≥ 30% improvement in TJC68
- ≥ 30% improvement in SJC66
- Improvement in PtGA of Disease Activity NRS
- Improvement in PGA of Disease Activity NRS

Four components are included in the PsARC criteria. Missing values for each component can occur due to data collection issues for a particular form, a missed visit or dropout from the study. In the case when PsARC responses cannot be determined for any visit date within a visit window, partially observed data from different visit dates within the visit window can be combined to determine PsARC responses for the visit window.

9.4.19 Ankylosing Spondylitis Disease Activity Score (ASDAS)

The Ankylosing Spondylitis Disease Activity Score (ASDAS) is a composite index to assess disease activity in AS. It combines five disease activity variables. Parameters used for the calculation of ASDAS:

1. Patient's assessment of total back pain (BASDAI Question 2)
2. PtGA of disease activity (0 - 10 NRS)
3. Peripheral pain/swelling (BASDAI Question 3)
4. Duration of morning stiffness (BASDAI Question 6)
5. High-sensitivity C-reactive protein (hs-CRP) in mg/L.

Calculation of ASDAS:

$$ASDAS_{hs\text{-}CRP} = 0.121 \times \text{total back pain} + 0.110 \times \text{patient global} + 0.073 \times \text{peripheral pain/swelling} + 0.058 \times \text{duration of morning stiffness} + 0.579 \times \text{Ln(hs-CRP+1).okie}$$

To calculate observed ASDAS scores, the observed component value will be calculated first. Then the components will be included in the calculation per the ASDAS formula. If

any observed component is missing in a window, then the observed ASDAS score will be missing.

ASDAS score is categorized in to the following ASDAS Disease Activity States:

- ASDAS Inactive Disease: ASDAS < 1.3
- ASDAS Moderate Disease: 1.3 ≤ ASDAS < 2.1
- ASDAS High Disease: 2.1 ≤ ASDAS ≤ 3.5
- ASDAS Very High Disease: ASDAS > 3.5

ASDAS Response categories are defined as follows:

- ASDAS Major Improvement (a change from baseline ≤ –2.000)
- ASDAS Clinically Important Improvement (a change from baseline ≤ –1.100)

9.4.20 FACIT-Fatigue Questionnaire (FACIT-F)

The FACIT Fatigue Questionnaire is a 13-item tool that measures an individual's level of fatigue during their usual daily activities over the past week. The level of fatigue is measured on a four point scale (4 = not at all fatigued to 0 = very much fatigued). The Fatigue scale ranges from 0 to 52, with higher scores indicating less fatigue.

Item score for each item is calculated by either subtracting from 4 or adding 0 depending on whether it is a reversal item or not. FACIT Fatigue is then calculated by adding up all item scores, multiplying by 13 and dividing by the number of items answered. If less than 7 items are answered, the scale will not be computed.

9.4.21 EuroQoL-5D (EQ-5D-5L)

EQ-5D measures 5 dimensions of health status (mobility, self-care, usual activities, pain/discomfort, and anxiety/depression) with 5 levels per dimension (no problems, slight problems, moderate problems, severe problems, and extreme problems corresponding to Level 1 to Level 5 respectively) and includes the EQ Visual Analogue Scale (EQ VAS).

The 5 dimensions of health status are converted into a single index value. The change from baseline of the index value and the EQ VAS will be analyzed and reported. UK scoring algorithm will be used.

9.4.22 Form SF-36v2

The 36-Item Short Form, Version 2 (SF-36v2) health survey consists of 36 general health questions. It has 2 components: physical and mental. For each component, a transformed summary score is calculated using 8 sub domains: physical functioning, role-physical, bodily pain, general health, vitality, social functioning, role-emotional, and mental health.

The coding and scoring for the SF-36 will use the software provided by QualityMetrics.

9.4.23 Work Productivity and Activity Impairment Questionnaire Psoriatic Arthritis (WPAI)

The Work Productivity and Activity Impairment Questionnaire: Psoriatic Arthritis, V2.0 (WPAI) measures the effect of overall health and specific symptoms on productivity at work and outside of work. It consists of 6 questions. A lower WPAI score indicates an improvement. The WPAI is collected at the designated study visits listed in the protocol. The WPAI coding and scoring methods are described in the following:

The 6 measures will be derived based on the responses from the 6 questions. The 4 main impairment scores (S1 to S4) are expressed as *percent impairment* based on the 6 questions.

Scores:

S0. Employment: defined below in missing data handling conventions

S1. Absenteeism: Percent work time missed due to PsA:

$$100 \times \left[\frac{Q2}{Q2 + Q4} \right]$$

S2. Presenteeism: Percent impairment while working due to PsA:

$$100\times\left[\frac{Q5}{10}\right]$$

S3. Percent overall work impairment due to PsA:

$$100\times\left[\frac{Q2}{Q2+Q4}+\left\{1-\frac{Q2}{Q2+Q4}\right\}\times\frac{Q5}{10}\right]$$

S4. Percent activity impairment due to PsA:

$$100\times\left[\frac{Q6}{10}\right]$$

S5. Did subject miss work (defined below). This is needed to derive the proportion of subjects who missed work.

Missing Data Handling Conventions

When calculating the WPAI: PsA scores, the following computational notes should be followed.

- Define Employment as a binary YES or NO variable where YES corresponds to "Employed" and NO corresponds to "Not Employed."
  - A subject will be considered "employed" at a given visit if Q1 = YES or Q2 > 0 or Q4 > 0.
  - A subject will be considered "unemployed" at a given visit if Q1 = NO and no positive hours recorded under Q2 and Q4 (i.e., if Q1 = NO AND Q2 ≤ 0 AND Q4 ≤ 0, then UNEMPLOYED).
  - Employment status for a subject will be considered "missing" at a given visit if Q1 = missing and no positive hours recorded under Q2 and Q4.
- If a subject is "unemployed" or employment status is "missing," then S1, S2, and S3 will be set to "missing."

- If Q2 = 0 and Q4 = 0 are missing then Q2/(Q2 + Q4) = missing (i.e., S1 = missing).
- If Q2 = 0 and Q4 = 0, then set S3 to missing.
- If Q2 is missing or Q4 is missing, then set S1 and S3 to missing.
- If Q4 = missing, then DO NOT set Q5 = missing.
- If Q5 is missing, then apply the following rules:
  - If Q2 > 0, Q4 = 0, and Q5 = missing, then S3 = 100%.
  - If Q2 = 0, Q4 > 0, and Q5 = missing, then S3 is missing.
  - If Q2 > 0, Q4 > 0, and Q5 = missing, then S3 is missing.
- Determine if a subject missed work (based on Q2) in order to analyze the proportion of subjects who missed work:
  - Create a binary (yes or no) "missed work" variable.
  - A subject will be considered as yes to missed work if Q2 is greater than 0.
  - If Q2 = missing, then MISSED WORK = missing.
  - If Q2 > 0, then MISSED WORK = "yes."
  - If Q2 = 0, then MISSED WORK = "no."

    Therefore, the proportion of subjects who missed work will be counted based on the number of subjects with MISSED WORK = YES.

9.4.24 Health Resource Utilization (HRU) Questionnaire

The HRU questionnaire contains three questions regarding health care utilization in the following categories: unscheduled health care professional visits, emergency room visits, and hospital admissions. The data gathered from the HRU questionnaire will be used to calculate the individual cumulative number of utilizations per time (e.g., subject-year) under observation in each variable (i.e., the number of unscheduled PsA-related health care professional visits, the number of emergency room visits, the number of hospital admissions and the total number of days in hospital) as follows:

- Time under observation for a subject will be defined as "date of last visit with non-missing HRU – date of baseline visit."

- The number of utilizations after baseline will be summed up for each subject.

To determine cumulative HRU over all subjects, the ratio of the total number of utilizations (i.e., over all subjects) and the total time under observation (i.e., over all subjects) will be calculated across all subjects in each treatment group. HRU will be analyzed as observed only.

9.4.25 Modified PsA Sharp/van der Heijde Score (SHS)

Radiographic outcomes will be assessed and scored according to Sharp's method (van der Heijde modification for PsA) centrally by two qualified readers who will be blinded to site number, subject number, treatment allocation, time sequence and clinical response.

Calculation of the SHS Score

To obtain the total SHS score, scores for erosions and JSN in both the hands and feet will be added together.

The following considerations are applied before score summation:

- Osteolysis in the form of pencil-in-cup: Osteolysis of the proximal phalanx and the base of the distal phalanx resulting in a pencil like proximal phalanx covered by cup like base of the distal phalanx. Pencil-in-cup will be scored as "P" where applicable.
- Gross Osteolysis: Osteolysis of the phalanx resulting in a loss of the normal joint structure, usually accompanied by shortening of the length of the phalanx. Gross osteolysis will be scored as "G" where applicable.

For the purpose of summing total scores, "G" or "P" assigned by the IRC reviewers will be associated with a maximum score per feature (JSN and erosion score) per location in analyses of radiographic scores.

Not Visible and Surgically Modified Joints/Bones

If a joint or bone is not visible (e.g., poor film quality, missing imaging, severe misalignment, flexion deformity, dislocation) at the timepoint, the individual joint or bone will be coded as Not Visible (N). If radiographs at the timepoint show a joint or bone with surgical fusion, replacement (prosthesis), or amputation, then the joint or bone will be scored Surgically Modified (S).

For joints and bones with end stage disease, scores N and S should not be used.

The range of scores is summarized below.

Table 10. Range of Total SHS Score, Erosion Score and Joint Space Narrowing

	Hands	Feet	Total (Hands and Feet)
Erosion Score Range	0 – 200	0 – 120	0 – 320
Joint Space Narrowing Range	0 – 160	0 – 48	0 – 208
Total SHS Range for Erosion and JSN	**0 – 360**	**0 – 168**	**0 – 528**

Erosion Assessment

Erosions will be assessed in each hand (20 locations per hand) and foot (6 locations per foot). The locations assessed in the SHS method include:

Hands:

- 4 Distal inter-phalangeal joints (2 - 5)
- 5 Metacarpo-Phalangeal Joints (1 - 5)
- 4 Proximal Inter-Phalangeal Joints (2 - 5)
- Inter-Phalangeal Joint of the thumb
- Proximal first Metacarpal Bone
- Radius Bone

- Ulnar Bone
- Trapezium and Trapezoid (as one unit; multangular)
- Navicular Bone
- Lunate Bone

Feet:

- 5 Metatarso-phalangeal joints (1 - 5)
- Inter-phalangeal joint of the first toe

Joint Space Narrowing Assessment

Joint space narrowing (JSN) will be assessed in each hand (20 locations per hand) and foot (6 locations per foot). The locations assessed in the SHS method include:

- Hand Joints:
  - 4 Distal inter-phalangeal joints (2 - 5)
  - 4 Proximal inter-phalangeal joints (2 - 5)
  - 5 Metacarpo-phalangeal joints (1 - 5)
  - Interphalangeal Joint of thumb (IP)
  - 3 Carpo-metacarpal joints (3 - 5)
  - Radio-carpal joint
  - Multangular-navicular joint
  - Capitate-navicular-lunate joint
- Foot Joints:
  - 5 Metatarso-phalangeal joints
  - Inter-phalangeal joint of first toe

For each Joint and Bone assessed, scores range as follows:

- Erosions: 0 – 5 (hands/wrists) or 0 – 10 (feet) to characterize the extent of erosions (where 0 denotes no erosion).

- Joint Space Narrowing: 0 – 4 to characterize the extent of Joint Space Narrowing (JSN) (where 0 denotes no narrowing).

Erosion and JSN scores for each reader are calculated by taking the sum of the left and right joints as shown below.

$$\text{Erosion}_{\text{Reader}i} = \text{Erosion}_{\text{Left}} + \text{Erosion}_{\text{Right}}$$

$$\text{JSN}_{\text{Reader}i} = \text{JSN}_{\text{Left}} + \text{JSN}_{\text{Right}}$$

for i = 1, 2.

Thus, the maximum erosion score for the hands is 200. The maximum erosion score for the feet is 120. Thus, the total erosion score for hands/wrists and feet is 320.

The maximum score for JSN in all 40 hand joints is 160. The maximum score for JSN in all 12 feet joints is 48. Thus, the total JSN score for hand and feet is 208.

Since two independent readers evaluate each film, the mean score will be calculated for the two readers from the individual erosion and JSN scores as shown below:

$$\text{Joint Erosion} = \frac{\text{Erosion}_{\text{Reader1}} + \text{Erosion}_{\text{Reader2}}}{2}$$

$$\text{JSN} = \frac{\text{JSN}_{\text{Reader1}} + \text{JSN}_{\text{Reader2}}}{2}$$

The SHS for each reader is defined as the sum of the erosion and JSN scores:

$$\text{SHS}_{\text{Reader i}} = \text{Erosion}_{\text{Reader i}} + \text{JSN}_{\text{Reader i}} \text{ for i} = 1, 2.$$

The average modified SHS from the two readers will be used for all x-ray endpoint calculations.

$$SHS = \frac{SHS_{reader1} + SHS_{reader2}}{2}$$

Handling of Missing Joints in the SHS Derivation

If a score at any location/joint is missing, the method described below will be used for deriving SHS.

- If the score for a location/joint is missing at Baseline, this joint will not contribute to the calculation of SHS for this subject at any visit within the reading session (even if the score for this location/joint is available at post-baseline visits).
- If the score for a location/joint is missing at all post-baseline visits within a reading session, this joint will not contribute to the calculation of SHS for this subject at any visit within the reading session (even if the score for this location/joint is available at Baseline).
- If the score for a location/joint is available at Baseline and at least one post-baseline visit, missing scores for this joint at any other post-baseline visit will be imputed assuming no progression from the previous time point with available score.

Adjudication Process

Two reviewers will independently review the images. Adjudication will occur for all subjects with a discrepancy of $\geq \left| 8 \text{ units} \right|$ between the two reviewers' SHS change scores, in which case another reviewer, different from the reviewers who performed primary assessments, will make a third, independent assessment.

For the calculation of SHS, the score of two closest reads (out of the two primary reviewers and the adjudicator) will be used. In the case of equal distance, the average of the three reads will be used.

10.0 Safety Analysis

10.1 General Considerations

Safety analyses will be carried out using the Safety Analysis Set by "as treated" treatment groups. There are two sets of planned safety analysis: safety analysis by Week 24, and long-term safety analysis.

Safety Analysis through Week 24

Standard safety analysis by the "as treated" treatment groups of upadacitinib 15 mg QD, upadacitinib 30 mg QD, adalimumab 40 mg EOW and combined placebo groups will be performed on safety data up to Week 24. No protocol-defined treatment switching will occur prior to this time point.

The standard safety analyses will include reporting of adverse events (AEs), laboratory, and vital signs measurements. Frequency tables of subjects with treatment-emergent adverse events (TEAEs) by system organ class (SOC) and by preferred term (PT) as in the Medical Dictionary for Regulatory Activities (MedDRA) dictionary will be provided by as-treated treatment groups. Mean changes from baseline in all continuous laboratory parameters and vital signs variables at each visit will be summarized by "as treated" treatment groups. Frequency tables of subjects meeting criteria for potentially clinically significant vital sign values and for potentially clinically significant laboratory values will be provided by as-treated treatment groups. Missing safety data will not be imputed. Treatment comparison of upadacitinib versus placebo and ADA will be provided for key safety endpoints including overview of AE, overview of AE of special interest, change from baseline in selected laboratory parameters, and potentially clinically significant laboratory values. Treatment difference and corresponding 95% CI will be provided.

Long-Term Safety Analysis

Long-term safety analyses that account for protocol-defined treatment switching include reporting of AE rate adjusted by cumulative exposure, descriptive summary in laboratory

parameters and vital sign variables by visit, and rate of potentially clinically significant laboratory and vital signs values. The treatment-emergent adverse event (TEAE) rate per 100 patient-years of exposure will be presented by actual treatment received at the time of AE (as described in Section 10.2.2). Listing of subjects with TEAEs by SOC and PT will be provided. Summary statistics for laboratory parameters and vital signs variables at each visit will be presented by "as treated" treatment group sequences defined below. Frequency tables and listings of subjects meeting criteria for potentially clinically significant vital sign values and for potentially clinically significant laboratory values will be provided by actual treatment received at the time of event. Missing safety data will not be imputed.

"As treated" treatment group sequences are defined as follows:

1. Placebo → Upadacitinib 15 mg QD
2. Placebo → Upadacitinib 30 mg QD
3. Upadacitinib 15 mg QD
4. Upadacitinib 30 mg QD
5. Adalimumab 40 mg EOW

Starting with Amendment 6.01 (for Japan) and Amendment 7 and 7.01(for global), subjects receiving upadacitinib 30 mg QD will be switched to upadacitinib 15 mg QD. For these subjects, adverse events and exposure to upadacitinib 30 mg QD will be censored at the time of dose switch; subsequent adverse events and exposure starting the day of first dose of upadacitinib 15 mg QD will be summarized under separate groups, as described in Section 8.2.

Assessment of potentially clinically significant laboratory and vital sign values will be based on the same treatment groups as for adverse events. The baseline value for PCS determination will be the last non-missing value collected before the first dose of UPA (regardless of dose).

For descriptive summary by visit or mean change from baseline in laboratory values and vital signs, exposure sequences 2 and 4 (i.e., subjects who switched from upadacitinib 30 mg QD to 15 mg QD) will be censored at the time of dose switch. Descriptive summary by visit or mean change from baseline for laboratory values and vital signs collected after dose switch will be summarized separately by visits post switch. Note for analysis of mean change from baseline in laboratory values, the study baseline is used.

Treatment comparisons of upadacitinib versus ADA will be provided for key safety endpoints including overview of AE, overview of AESI, selected lab parameters, and potentially clinically significant laboratory values. Treatment difference and corresponding 95% CI will be provided.

10.2 Analysis of Adverse Events

A treatment-emergent Adverse Event (TEAE) is defined as an adverse event with an onset date that is after the first dose of study drug, and no more than 30 days after the last dose of upadacitinib and placebo, and 70 days after the last dose of adalimumab.

Events where the onset date is the same as the study drug start date are assumed to be treatment-emergent, unless the study drug start time and the adverse event start time are collected and the adverse event start time is prior to the study drug start time. If an incomplete onset date was collected for an adverse event, the event will be assumed to be treatment-emergent unless there is other evidence that confirms that the event was not treatment-emergent (e.g., the event end date was prior to the study drug start date).

Adverse event data will be presented by SOCs and PTs using MedDRA Version 19.1 or most up to date version, which will be sorted in alphabetical order by SOC and PT.

10.2.1 Analysis of Adverse Events Prior to Protocol-Defined Treatment Switching

10.2.1.1 Adverse Events Overview

The number and percentage of subjects experiencing TEAEs will be summarized by treatment groups for the following AE categories.

- All TEAEs
- Treatment-emergent serious adverse events (SAEs)
- Treatment-emergent severe adverse events
- TEAEs with a reasonable possibility of being related to study drug
- TEAEs leading to discontinuation of study drug
- TEAE leading to death
- Deaths (includes all deaths – treatment-emergent and non-treatment-emergent)

In the AE overview summary, any event with an unknown severity will be considered as severe and any AE with an unknown relationship will be considered as having "reasonable possibility" of being related to study drug. Additional AE summaries may be considered for tabulation/summary based on recommendations from Clinical and Safety as deemed appropriate.

10.2.1.2 Adverse Events by System Organ Class and Preferred Term

The number and percentage of subjects experiencing adverse events a will be tabulated by SOC and MedDRA PT by treatment groups. The SOCs will be presented in alphabetical order, and the PTs will be presented in alphabetical order within each SOC.

The following summaries of adverse events will be generated:

- All TEAEs
- Treatment-emergent serious adverse events (SAEs)
- Treatment-emergent severe adverse events
- TEAEs with a reasonable possibility of being related to study drug

- TEAEs leading to discontinuation of study drug
- TEAE leading to death

Subjects reporting more than one adverse event for a given MedDRA preferred term will be counted only once for that term (most severe incident for the severity tables and most related incident for the relationship tables). Subjects reporting more than one type of adverse event within a SOC will be counted only once for that SOC. Subjects reporting more than one type of adverse event will be counted only once in the overall total.

10.2.1.3 TEAEs by Maximum Severity

TEAEs will also be summarized by maximum severity by treatment groups. If a subject has an AE with an unknown severity, then the subject will be counted in the severity category of unknown, even if the subject has another occurrence of the same event with a severity present. The only exception is if the subject has another occurrence of the same AE with the highest CTCAE grade; in this case, the subject will be counted under the highest CTCAE grade for the term in question.

10.2.1.4 TEAEs by Relationship

TEAEs will also be summarized by relationship to upadacitinib, adalimumab and placebo, as assessed by the investigator, by treatment groups. If a subject has a TEAE with an unknown relationship, then the subject will be counted in the relationship category of "unknown," even if the subject has another occurrence of the same event with a relationship present. The only exception is if the subject has another occurrence of the same TEAE with a relationship assessment of "reasonable possibility." In this case, the subject will be counted under the "reasonable possibility" category.

10.2.1.5 Frequent (≥ 2%) Adverse Events and Reasonably Possibly Related Adverse Events by System Organ Class and Preferred Term

TEAEs and reasonably possibly related AEs occurring for more than 2% of the subjects in any of the treatment groups will be summarized by SOC and PT separately.

10.2.1.6 Adverse Events of Special Interest

The Adverse Events of Special Interest (AESI) categories will be summarized and presented by treatment groups in overview as well as using SOC and PT. The AESI categories will be identified by the following search criteria per Standard MedDRA Queries (SMQs)/Company MedDRA Queries (CMQs) in Table 11 below. Adjudicated cardiovascular events will be summarized and presented by treatment group using the CAC adjudicated categories.

Table 11. AESI for Upadacitinib with SMQs/CMQs/PTs Searches

AESI	Type of MedDRA Query	Broad or Narrow Search	SMQ/CMQ Search Criteria
Serious Infections	CMQ		"Infections" – Subset for SAEs
Opportunistic Infection excluding Tuberculosis and Herpes Zoster	CMQ		"Opportunistic Infection excluding Tuberculosis and Herpes Zoster"
Possible Malignancy	SMQ	Narrow	"Malignancies"
Malignancy	SMQ	Narrow	"Malignant tumours"
Non-Melanoma Skin Cancer (NMSC)	SMQ	Narrow	Skin Malignant tumours (Narrow SMQ) removing Melanoma CMQ
Malignancy excluding NMSC			"Malignant tumours" SMQ (Narrow) removing NMSC output
Lymphoma	SMQ	Broad	"Malignant Lymphomas"
Hepatic Disorder	SMQ	Narrow	"Drug Related Hepatic Disorders"
Adjudicated Gastrointestinal Perforations	Based on adjudicated results (the identification of events to be adjudicated are described in the GI Perforation charter)		
Anemia	CMQ		"Non-Hemolytic and Non-Aplastic Anemias"
Neutropenia	CMQ		"Hematological Toxicity – Neutropenia"
Lymphopenia	CMQ		"Hematological Toxicity – Lymphopenia"
Herpes Zoster	CMQ		"Herpes Zoster"
Creatine Phosphokinase (CPK) Elevation	PT		Search only for the PT of "Blood creatine phosphokinase increased"
Renal Dysfunction	SMQ	Narrow	"Acute Renal, Failure"
Active Tuberculosis	CMQ		"Active Tuberculosis"

Table 11. AESI for Upadacitinib with SMQs/CMQs/PTs Searches (Continued)

AESI	Type of MedDRA Query	Broad or Narrow Search	SMQ/CMQ Search Criteria
Adjudicated cardiovascular events	Output from CAC		
MACE*			
Cardiovascular Death			
Non-fatal Myocardial Infarction			
Non-fatal Stroke			
Other Adjudicated Cardiovascular Events			
Undetermined/Unknown Cause of Deaths			
Adjudicated Thrombotic Events	Output from CAC		
VTE**			
Deep Vein Thrombosis			
Pulmonary Embolism			
Other Venous Thrombosis			
Arterial Thromboembolic Events (non-cardiac, non-neurologic)			

CAC = Cardiovascular Adjudication Committee; CMQ = company MedDRA query; PT = preferred term; SMQ = standard MedDRA query

* MACE; Major Adverse Cardiovascular Events, defined as cardiovascular death, non-fatal myocardial infarction and non-fatal stroke.

** VTE: Venous thromboembolic events, defined as deep vein thrombosis (DVT) and pulmonary embolism (PE) (fatal and non-fatal).

Additional AEs may be considered for tabulation/summary based on recommendations from Clinical and Safety as deemed appropriate.

10.2.2 Analysis of Long-Term Adverse Event Rates

Long-term adverse events will be analyzed using event rates adjusted by cumulative exposure and will be based on the actual treatment received at the time of AE occurrence. For subjects switched from upadacitinib 30 mg QD to upadacitinib 15 mg QD, adverse events and exposure to upadacitinib 30 mg QD will be censored at the time of dose switch; subsequent adverse events and exposure starting the day of first dose of upadacitinib 15 mg QD will be summarized under separate groups. The detailed treatment groups including the handling of dose switch, are described in Section 8.1.

Exposure-Adjusted Event Rate (EAER)

To adjust for potentially different follow-up time between treatment groups, EAER will be provided. For the purpose of event rate calculation, the numerator will be the total number of AEs reported for the event (i.e., a subject can contribute more than one event to the numerator) and the denominator will be the total exposure time among subjects under the treatment group. The number of AEs reported (numerator), the total number of years of study drug exposure (denominator), and the exposure-adjusted AE event rate per 100 patient-years, calculated as ([numerator/denominator])*100, will be presented for each treatment group. The EAER will be the main approach to evaluate AEs in the long-term analysis.

The exposure adjusted incidence rate (censored at the time of first event) may be conducted for selected AESI endpoints as appropriate for long-term analysis.

10.2.2.1 Overview of Adverse Events Rates per 100 Patient-Years of Study Drug Exposure

An overview of AEs per 100 patient-years of study exposure will be presented by treatment group for the following AE categories.

- All TEAEs
- Treatment-emergent serious adverse events (SAEs)
- Treatment-emergent severe adverse events

- TEAEs reasonably possibly related to study drug
- TEAEs of special interest
- TEAEs leading to discontinuation of study drug
- TEAE leading to death
- Deaths (includes all deaths – treatment-emergent and non-treatment-emergent)

In the AE overview summary, any event with an unknown severity will be considered as severe and any AE with an unknown relationship will be considered as having "reasonable possibility" of being related to study drug. Additional AEs may be considered for tabulation/summary based on recommendations from Clinical and Safety as deemed appropriate.

10.2.2.2 Adverse Events Rates per 100 Patient-Years of Study Drug Exposure by SOC and PT

For each treatment group, the TEAE rate per 100 patient-years of exposure will be calculated overall, for each SOC and each PT, for each of the following events:

- All TEAEs
- Treatment-emergent serious adverse events (SAEs)
- Treatment-emergent severe adverse events
- TEAEs with a reasonable possibility of being related to study drug
- TEAEs leading to discontinuation of study drug
- TEAE leading to death

10.2.2.3 Adverse Events of Special Interest Rates per 100 Patient-Years of Study Drug Exposure

The Adverse Events of Special Interest (AESI) categories will be summarized and presented for each treatment group in overview as well as using SOC and PT. The AESI categories will be identified per Standard MedDRA Queries (SMQs)/Company MedDRA

Queries (CMQs). Adjudicated cardiovascular events will be summarized and presented by treatment group using the CAC adjudicated categories.

10.2.2.4 Listing of Serious Adverse Events (Including Deaths) and Adverse Events Leading to Study Drug Discontinuation

All serious adverse events (SAEs), deaths, and adverse events leading to discontinuation of study drug will be listed.

10.3 Analysis of Laboratory Data

10.3.1 Variables and Units

All laboratory parameters to be collected in this study are listed below. Laboratory parameters will be reported using the standard international (SI) units.

Table 12. List of Laboratory Variables

Laboratory Variables
Hematology
White Blood Cell (WBC) Count
Red Blood Cell (RBC) Count
Hemoglobin
Hematocrit
Platelets count
Neutrophils
Basophils
Eosinophils
Lymphocytes
Monocytes
Bands
Chemistry
Total Bilirubin
Alkaline Phosphatase (ALP)
Aspartate aminotransferase (AST)
Alanine aminotransferase (ALT)
Total Protein
Albumin
Glucose
Triglycerides
Blood Urea Nitrogen (BUN)
Creatinine
Sodium
Potassium
Calcium
Inorganic Phosphorus
Creatine Phosphokinase (CPK)
Chloride
Bicarbonate

Table 12. List of Laboratory Variables (Continued)

Laboratory Variables
Chemistry (Continued)
Cholesterol
LDL cholesterol
HDL cholesterol
LDL/HDL ratio
Cholesterol/HDL ratio
Urinalysis
Specific Gravity
pH
Protein
Glucose
Ketones
Blood
Microscopic Examination (if needed)
Urobilinogen
Bilirubin
Leukocytes
Nitrites
Other
hs-CRP
ESR

10.3.2 Analysis of Laboratory Data by Week 24

The laboratory data will be summarized by the "as treated" treatment groups (Upadacitinib 15 mg QD, Upadacitinib 30 mg QD, ADA 40 mg EOW and combined placebo groups).

10.3.2.1 Assessment of Clinical Laboratory Variables

Analyses of hematology, chemistry, and urinalysis variables which are measured longitudinally will be performed by visits and by treatment groups. For analysis at each

visit, the following summary statistics of visit values will be presented for each treatment group: sample size, mean, standard deviation, minimum, median, and maximum.

An ANOVA model with treatment as a factor will be used to compare change from baseline between different treatment groups for selected laboratory parameters. Mean difference from placebo/ADA and associated 95% CIs will be presented. The analysis applies to the following laboratory parameters of clinical interest: hemoglobin, platelets, lymphocytes, neutrophils, creatinine, ALT, AST, creatine phosphokinase (CPK), LDL, HDL, total cholesterol/HDL-cholesterol, LDL-cholesterol/HDL-cholesterol, and total cholesterol.

10.3.2.2 Assessment of Shift from Baseline in Clinical Laboratory Variables

The baseline and post-baseline laboratory observations will be categorized as Grade 0, Grade 1, Grade 2, Grade 3, and Grade 4 and shifts from baseline grade to worst on-therapy grade will be summarized. Toxicity grading scale is based on National Cancer Institute Common Toxicity Criteria (NCI CTC) AE version 4.03. Shift tables from Baseline according to the grades will be provided for laboratory variables.

For LDL cholesterol, HDL cholesterol, total cholesterol, and triglycerides, the following categories according to National Cholesterol Education Program (NCEP) Adult Treatment Panel III (ATP III) guidelines will be used.

- LDL cholesterol (< 3.36, ≥ 3.36 and < 4.14, ≥ 4.14 mmol/L)
- HDL cholesterol (< 1.03, ≥ 1.03 mmol/L)
- Total cholesterol (< 5.17, ≥ 5.17 and < 6.21, ≥ 6.21 mmol/L)
- Triglycerides (< 1.69, ≥ 1.69 and < 2.26, ≥ 2.26 mmol/L)

Note that the minimum/maximum category is used, rather than the category of the minimum/maximum value. The two may be different due to variation in the reference range.

No statistical tests will be performed for this analysis.

10.3.2.3 Assessment of Potentially Clinical Significant Laboratory Values

The criteria for potentially clinically significant laboratory values will be determined by CTCAE criteria of Grade 2, Grade 3, Grade 4 and ≥ Grade 3(if applicable), with a grade worsening compared to baseline. Toxicity grading scale is based on National Cancer Institute Common Toxicity Criteria (NCI CTC) AE version 4.03. The number and percentage of subjects meeting the criteria for potentially clinically significant laboratory values will be summarized by treatment groups.

10.3.2.4 Assessment of Liver Elevations

According to FDA's Guidance for Industry "Drug-Induced Liver Injury: Premarketing clinical evaluation" (July 2009), when aminotransferase (AT) abnormalities indicating hepatocellular injury are accompanied by evidence of impaired hepatic function (bilirubin elevation > 2 × ULN), in the absence of evidence of biliary obstruction (i.e., significant elevation of ALP) or some other explanation of the injury (e.g., viral hepatitis, alcohol hepatitis), the combined finding (i.e., Hy's Law cases) represents a signal of a potential for the drug to cause severe DILI.

For the purpose of assessing for potential Hy's law cases, the frequencies and percentages of subjects with post baseline liver specific function test values that meet the following criteria of potential clinical interest will be summarized by treatment groups:

- ALT ≥ 3 × ULN
- ALT ≥ 5 × ULN
- ALT ≥ 10 × ULN
- ALT ≥ 20 × ULN
- AST ≥ 3 × ULN
- AST ≥ 5 × ULN
- AST ≥ 10 × ULN

- AST ≥ 20 × ULN
- TBL ≥ 2 × ULN
- Alkaline phosphatase ≥ 1.5 × ULN
- ALT and/or AST ≥ 3 × ULN and concurrent TBL ≥ 1.5 × ULN
- ALT and/or AST ≥ 3 × ULN and concurrent TBL ≥ 2 × ULN

10.3.3 Analysis of Long-Term Laboratory Data

10.3.3.1 Assessment of Clinical Laboratory Variables

Analyses of hematology, chemistry, and urinalysis variables which are measured longitudinally will be performed by visits and by treatment group sequences as described in Section 10.1. For each analysis, the following summary statistics of visit values will be presented for each treatment group sequence: sample size, mean, standard deviation, minimum, median and maximum.

Analyses will be performed for change from baseline in hemoglobin, lymphocytes, neutrophils, creatinine, and creatine phosphokinase (CPK).

Descriptive summary by visit or mean change from baseline in laboratory values for subjects who switched from upadacitinib 30 mg QD to 15 mg QD will be censored at the time of dose switch, and descriptive summary by visit or mean change from baseline for values post switch will be summarized separately, as described in Section 10.1.

10.3.3.2 Assessment of Potentially Clinical Significant Laboratory Values

Long-term laboratory data will be summarized based on the number and percentage of subjects meeting the criteria for potentially clinical significant laboratory values and by the actual treatment received at the time of the event occurrence. The treatment groups for summarizing potentially clinically significant laboratory values are the same as the ones for long-term AE analysis as described in Section 8.1.

A subject can be counted under different treatment groups if he/she was on placebo from baseline to Week 24 and then started upadacitinib 15 or 30 mg at Week 24 and experienced potentially clinical significant laboratory values under different treatment groups.

In the evaluation of potentially clinically significant laboratory values, the baseline value will be determined by the last non-missing measurement recorded on or before the date of the first dose of study drug in the corresponding treatment group (which may be different than the first dose of study drug received in the study). For example, for a subject who started on placebo and switched to upadacitinib 15 mg QD at Week 24, lab values under upadacitinib 15 mg QD exposure would be evaluated against the baseline value defined as the last non-missing measurement recorded on or before the date of the first dose of upadacitinib 15 mg QD. For subjects who switched from upadacitinib 30 mg QD to upadacitinib 15 mg QD, lab values under upadacitinib 15 mg QD exposure would still be evaluated against the baseline defined based on the first dose of Upadacitinib 30 mg QD.

A listing of all subjects with any laboratory determination meeting CTCAE criteria of Grade 3 or above with a grade worsening compared to baseline will be provided. For each of these subjects, the whole course of the respective parameter will be listed.

10.3.3.3 Assessment of Liver Elevations

The frequencies and percentages of subjects with post-baseline liver-specific function test values that meet the following criteria of potential clinical interest will be summarized by the actual treatment received at the time of the event occurrence. The treatment groups for summarizing liver elevations are the same as the ones for long-term AE analysis as described in Section 8.1.

A subject can be counted under different treatment groups if he/she was on placebo from baseline to Week 24 and then started upadacitinib 15 or 30 mg at Week 24 and experienced potentially clinical significant laboratory values under different treatment groups.

A listing of potentially clinically significant liver elevations based on criteria specified above will be provided. For each of these subjects, the whole course of the respective parameter will be listed.

10.4 Analysis of Vital Signs

10.4.1 Variables and Criteria Defining Abnormality

Vital sign variables include sitting systolic blood pressure, sitting diastolic blood pressure, pulse rate, and weight. The criteria for potentially clinically significant vital sign findings are presented in Table 13.

Table 13. Criteria for Potentially Clinically Significant Vital Sign Findings

Vital Sign	Category	Criteria for Potential Clinically Significant Vital Signs
Systolic blood pressure	Low	Value ≤ 90 mmHg and decrease ≥ 20 mmHg from Baseline
	High	Value ≥ 160 mmHg and increase ≥ 20 mmHg from Baseline
Diastolic blood pressure	Low	Value ≤ 50 mmHg and decrease ≥ 10 mmHg from Baseline
	High	Value ≥ 100 mmHg and increase ≥ 10 mmHg from Baseline
Weight	High	> 7% increase from baseline
	Low	> 7% decrease from baseline

10.4.2 Analysis of Vital Signs by Week 24

Analyses of vital sign variables which are measured longitudinally will be performed by visits and by the treatment groups of upadacitinib 15 mg QD, upadacitinib 30 mg QD, adalimumab 40 mg EOW, and the combined placebo group. For each analysis, the following summary statistics will be presented for each treatment group: sample size, mean, standard deviation, minimum, median, and maximum.

The number and percentage of subjects meeting the criteria for potentially clinically significant vital sign values will be summarized by treatment group.

10.4.3 Long-Term Analysis of Vital Signs

Analyses of vital signs variables which are measured longitudinally will be performed by visits and by treatment group sequences as described in Section 10.1. For each analysis, the following summary statistics will be presented for each treatment group: sample size, mean, standard deviation, minimum, median, and maximum.

Long-Term Vital Sign will also be summarized based on the number and percentage of subjects meeting the criteria for potentially clinical significant vital sign values and by the actual treatment received at the time of the event occurrence. The treatment groups are the same as the ones for long-term AE analysis as described in Section 10.2.2. A subject can be counted under different treatment groups if he/she was on placebo from baseline to Week 24 and then started upadacitinib 15 or 30 mg at Week 24 and experienced potentially clinical significant laboratory values under different treatment groups. In the evaluation of potentially clinically significant vital sign values, the baseline value will be determined by the last non-missing measurement recorded on or before the date of the first dose of study drug in the corresponding treatment group, similarly as described in Section 10.3.3.2.

Descriptive summary by visit in vital signs for subjects who switched from upadacitinib 30 mg QD to 15 mg QD will be censored at the time of dose switch, and descriptive summary by visit for values post switch will be summarized separately, as described in Section 10.1. The treatment groups for summarizing PCS vital sign values are the same as the ones for long-term AE analysis as described in Section 8.1. For the purpose of evaluating PCS vital sign values, for subjects who switched from upadacitinib 30 mg QD to upadacitinib 15 mg QD, baseline is defined based on the first dose of Upadacitinib 30 mg QD, similarly as described in Section 10.3.3.2.

A listing of all subjects with any vital sign values meeting the criteria for potentially clinically significant vital signs will be provided. For each of these subjects, the whole course of the respective parameter will be listed.

11.0 Summary of Changes

11.1 Summary of Changes between the Previous Version and the Current Version

The primary efficacy analysis has been completed under the SAP version 3.0 and SPP 1.0, Period 1 reporting has been completed under the SAP version 3.0 and SPP version 2.0. The current SAP update applies only to future reporting of long-term analysis.

1. Updated throughout the SAP to align with Protocol Amendment 6.01 and Amendment 7 and 7.01, including handling of dose switch from upadacitinib 30 mg QD to 15 mg QD (hereafter referred to as "dose switch").
    a. Updated Section 7.0 and Section 8.1 to provide analysis details that incorporate dose switch.
    b. Updated Section 9.1.1 to summarize the analysis changes that incorporate dose switch for long-term efficacy analysis.
    c. Updated Section 10.1, Section 10.2.2, Section 10.3.3.1, Section 10.3.3.2, Section 10.3.3.3 and Section 10.4.3 to incorporate handling of dose switch for long-term safety analysis.
2. Updated Section 9.1.2 and Section 9.3 for long-term binary endpoint analysis for Week 104 and future analyses.
    a. Updated Section 9.1.2 to describe the missing data handling approach for the additional long-term efficacy analysis using MMRM/GLMM models.
    b. Updated Section 9.3 to add NRI and/or GLMM for long-term non-radiographic binary endpoint analysis (Week 104 and future).
    c. Updated Section 9.3 to add MMRM based on AO data for long-term continuous endpoint analysis, including radiographic continuous endpoints (Week 104 and future).

3. Updated language in Section 10.1, Section 10.2.1.1, Section 10.2.2.1, Section 10.2.2.3 to add clarity for safety analysis.
4. Updated Section 10.2.1.6 and Section 10.4.1 to align with the latest upadacitinib AESI definitions in PSSAP V4.0.
5. Added Appendix E to describe analysis accounting for impact of COVID-19 pandemic.

12.0 Version History

Table 14. SAP Version History Summary

Version	Date	Summary
1.0	12 Apr 2018	Original version submitted to FDA for feedback.
2.0	12 Dec 2018	Revised by adding estimand for the primary and ranked secondary endpoints based on FDA's feedback.
3.0	12 Dec 2019	This version was approved and submitted to FDA prior to Week 24 DBL and unblinding. Endpoints were revised as part of Protocol Amendment 4; Revised tipping point analysis for the primary endpoints based on FDA's feedback.
4.0	08 Jul 2021	Incorporating update 1) in the study duration from 3 to 5 years in protocol amendment 6.0; 2) in statistical analysis to handle the missing data due to COVID-19; 3) in wording changes in PSSAP; 4) in handling of dose switch from upadacitinib 30 mg QD to 15 mg QD; 5) Add additional long-term analysis method for Week 104 and future visits.

13.0 References

1. Bretz F, Maurer W, Brannath W, et al. A graphical approach to sequentially rejective multiple test procedures. Stat Med. 2009;28(4):586-604.
2. Greenland S, Rothman KJ, Lash TL. Introduction to stratified analysis. In: Rothman KJ, Greeland S, Lash TL, editors. Modern Epidemiology. 3rd Ed. Philaphdelphia: Lippincott Williams & Wilkins; 2008.

3. Koch GG. Comments on 'current issues in non-inferiority trials'. Stat Med. 2008;27(3):333-42.
4. Carpenter JR, Kenward MG. Multiple Imputation and Its Application. New York: John Wiley & Sons; 2013.
5. Ratitch B, Lipkovich I, O'Kelly M. Combining analysis results from multiply imputed categorical data. In: Conference Proceedings from PharmaSUG 2013. May 12-15, 2013; Chicago, IL. Paper SP03.
6. Liu GF, Wang J, Liu K, et al. Confidence intervals for an exposure adjusted incidence rate difference with application to clinical trials. Stat Med. 2006;25(8):1275-86.

Appendix A. SAS Procedure for Mantel-Haenszel Test

A SAS procedure, the Proc Freq, will be used to compute the statistics using the Mantel-Haenszel method.[2] Risk difference, 95% CI and p value will be provided through the program.

SAS code example:

```
title 'Placebo vs ABT-494 30 mg QD';
proc freq data= wk12_nri;
where TRTP='Placebo'|TRTP='ABT-494 30 mg QD';
table TRTP*aval/ nopercent nocol chisq riskdiff(cl=wald)
alpha=0.05;
run;


title2 'For p-values';
proc freq data=wk12_nri;
where TRTP='Placebo'|TRTP='ABT-494 30 mg QD';
tables DMARDBL*TRTP*aval/ cmh;
run;
```

Appendix B. Tipping Point Analysis for ACR20 at Week 12

To assess the robustness of the primary analysis using NRI data handling, tipping point analysis is conducted on the primary endpoint ACR20. The analysis is conducted on the FAS using As Observed data handling.

The tipping point analysis will be performed by multiple imputations using logistic regression, allowing the imputed ACR response rate to systematically vary from 0% to 100% in both upadacitinib and placebo, respectively. This will be accomplished by modifying the predicted probabilities for the responses through shifting the log odds ratios,[4] then directly sampling the missing ACR20 response from the Bernoulli distribution with the modified probabilities.

For each pair of shift parameters, the same CMH method used for the primary endpoint will be performed on each of the multiple imputed datasets to obtain the results for each comparison of the upadacitinib treatment group versus the placebo group. Because chi-square distribution of CMH test statistic is skewed from the normal distribution, to combine the results from the CMH test using Rubin's method, a transformation will be conducted to normalize the CMH statistic.[5] These transformed results will be then aggregated using Rubin's method to get *P*-values.

Appendix C. Three Arm Approach to Non-Inferiority (NI) Testing

Let parameters T, R and P represent the ACR20 response rates for an upadacitinib dose group, the adalimumab group, and the combined placebo group respectively and let *n* be the sample size per group (as all groups have an equal sample size). Let α denote the type 1 error rate and β denote the type 2 error rate (so 1-β is the test power). The null and alternative hypotheses of non-inferiority testing have the form:

$$H_0: T - R \leq -(1-\phi)(R-P) \;\; vs. \;\; H_1: T - R > -(1-\phi)(R-P).$$

ϕ is a pre-specified fraction of reference drug effect to be retained by test drug, and it is set to be 50% in this NI comparison. Using the three-arm approach, both $T - R$ and $R - P$ are estimated from the current clinical trial. The three-arm approach test statistic has the following form:

$$Z = \frac{\bar{Y}_T - \bar{Y}_R + (1-\phi)(\bar{Y}_R - \bar{Y}_P)}{\sqrt{\frac{\bar{Y}_T(1-\bar{Y}_T)}{n} + \frac{\phi^2 \bar{Y}_R(1-\bar{Y}_R)}{n} + \frac{(1-\phi)^2 \bar{Y}_P(1-\bar{Y}_P)}{n}}},$$

where $\bar{Y}_T, \bar{Y}_R$ and $\bar{Y}_P$ are the sample ACR20 response rates for the upadacitinib dose group, the adalimumab group, and the combined placebo group respectively. The null hypothesis is rejected at a two-sided α level if $Z > Z_{\alpha/2}$, where $Z_{\alpha/2}$ is the upper $\alpha/2$ percentile of the standard normal distribution.

Appendix D. Details of Supplementary Analyses for Ranked Secondary Continuous Endpoints

Tipping Point Analysis for Key Secondary Non-Radiographic Continuous Endpoints

To assess the impact of potential departures from the missing-at-random assumption, tipping point analyses are conducted as a sensitivity check for change from baseline in key secondary non-radiographic continuous endpoints at the protocol defined primary time point.

The tipping point analyses are two-dimensional, i.e., assumptions about the missing outcomes on the upadacitinib treatment groups and the placebo group are allowed to vary independently. In addition, the focus is on scenarios where missing outcomes on upadacitinib are worse than the imputed values on upadacitinib, while missing outcomes on placebo are better than the imputed values on placebo. Missing values are first imputed via MI under MAR (where the MI imputation is performed upon AO data), and then a shift parameter is applied to the imputed values (a different shift parameter may be specified for each treatment group). This is implemented by PROC MI using the MNAR statement.

More specifically, PROC MI using the fully conditional specification (FCS) method will be performed to impute any missing value and therefore does not rely on a monotone missing pattern. In fact, the FCS method can handle any arbitrary missing data pattern. Treatment is included in the FCS imputation model to enable sampling conditional on treatment groups. Additionally, the imputation model includes stratification factor of current use of non-biologic DMARD (yes/no), gender, race (white vs. non-white), age, baseline BMI, geographic regions, duration of PsA diagnosis and the baseline value of the endpoints of interest, as well as longitudinal response observed at any other visits. The SAS code example is as follows:

For a given pair of shift parameters, the SAS code example is as follows:

```
PROC MI DATA=DATA_WIDE OUT=DATA_WIDE_BOUNDED NIMPUTE=20
SEED=12345;

  CLASS TRTP &COVCAT;
  FCS;
  By trtp;
  VAR &COVCON &COVCAT WEEK_2 WEEK_4 WEEK_8 WEEK_12;
MNAR ADJUST (WEEK_12 / SHIFT=&SJ1
ADJUSTOBS=(TRTP='PLACEBO'));
MNAR ADJUST (WEEK_12 / SHIFT=&SJ2 ADJUSTOBS=(TRTP='UPA
15MG'));
RUN;
```

Note: The input dataset is in wide format. TRTP denotes the treatment group. &COVCON denotes baseline continuous covariates including baseline age, baseline BMI, duration of PsA diagnosis and baseline value of the endpoint of interest. &COVCAT denotes categorical covariates including current use of non-biologic DMARD, gender, race, and regions. WEEK_2, WEEK_4, WEEK_8, WEEK_12 denote the observed values at each visit. &SJ1and &SJ2 denote the shift parameters for the placebo group and upadacitinib 15 MG group respectively.

In cases where the shifted values are smaller than the minimum or larger than maximum value of the endpoint, (i.e., out of range), the minimum or maximum value of the endpoint is used in further analysis steps. For each pair of shift parameters, the SAS procedure PROC MIXED is used for ANCOVA analysis with Huber-White sandwich errors which includes the fixed effects of treatment, the stratification factor of current DMARD use and the continuous fixed covariate of baseline measurement on each of the multiple imputed datasets to obtain the results for each upadacitinib treatment group versus the placebo group comparison. These results will be aggregated using Rubin's method to get p-values.

If one pair of shift parameters are found to just reverse the study conclusion, in terms of p-value larger than 0.05 (the original p-value < 0.05), then the shift parameters are identified as the tipping point. The results for a grid of shift parameter combinations are provided in tabular format.

The SAS code example of the data imputation step for tipping point analysis using MI is provided above. The SAS code example for the analysis and results combination step using PROC MIXED and PROC MIANALYZE is as follows:

```
PROC MIXED DATA=ALL EMPIRICAL;
```

```
    ODS OUTPUT LSMEANS=MIXEDLSMEANS DIFFS=DIFF;
    BY Shift1 Shift2 _IMPUTATION_;
    CLASS TRTP STRATA USUBJID;
    MODEL CHG = BASELINE TRTP STRATA / SOLUTION;
    LSMEANS TRTP / CL PDIFF DIFF;
    Repeated/subject = USUBJID;
RUN;

DATA DIFF1;
    SET DIFF;
    COMPARISON= TRTP||' VS '||LEFT(_TRTP);
RUN;

PROC SORT DATA=DIFF1; BY COMPARISON Shift1 Shift2 _IMPUTATION_; RUN;

PROC MIANALYZE DATA=DIFF1;
    ODS OUTPUT PARAMETERESTIMATES=GROUP_OUTPUT;
    BY comparison shift1 shift2;
    MODELEFFECTS ESTIMATE;
    STDERR STDERR;
RUN;

/* Comparison of "UPA 15 mg" vs. "Placebo" */
data miparm1;
set GROUP_OUTPUT;
if comparison='1            VS            2' then output miparm1;
run;
proc transpose data=miparm1(keep= Shift1 Shift2 Probt)
out=wide_miparm1;
by shift1;
id shift2;
var Probt;
run;
```

Note: The input dataset ALL includes all (# of shift1 parameters) * (# of shift2 parameters) * (# of imputations in MI) imputed datasets. TRTP denotes the treatment group and STRATA denotes the stratification factor used in analysis. CHG denotes the change from baseline value and BASELINE denotes the baseline value.

MI Analysis for Radiographic Continuous Endpoints

The MI analysis will impute missing data multiple times under appropriate random variation and thus generate multiple imputed "pseudo-complete" datasets. The imputation assumes of missing at random (MAR). ANCOVA will be performed on each of the multiple imputed datasets. The results will be aggregated across the multiple imputed datasets using Rubin's method. The SAS procedure PROC MI with the monotone

statement and REGRESSION method is used. Before applying PROC MI, MCMC is applied where necessary to augment data into monotone missing pattern. The SAS code example for PROC MI is as follows:

```
PROC MI DATA=DATA_WIDE OUT=OUTMI NIMPUTE=1 SEED=12345;


     Mcmc impute = monotone;
     VAR &COVCON base aval;
      By trtp;
RUN;


PROC MI DATA=DATA_WIDE OUT=OUTMI NIMPUTE=20 SEED=12345;

     CLASS &COVCAT;
     MONOTONE;
     VAR &COVCON &COVCAT base aval;
      By trtp;
RUN;
```

Note: The input dataset is in wide format. TRTP denotes the treatment group. &COVCON denotes baseline continuous covariates including baseline age, baseline BMI, duration of PsA diagnosis and baseline value of the endpoint of interest. &COVCAT denotes categorical covariates including current use of non-biologic DMARD, gender, race and regions. WEEK_16, WEEK_24 denote the observed values at each visit.

```
PROC MIXED DATA=ALL;
    ODS OUTPUT LSMEANS=MIXEDLSMEANS DIFFS=DIFF;
    BY _IMPUTATION_;
    CLASS TRTP STRATA;
    MODEL CHG = BASELINE TRTP STRATA / SOLUTION;
    LSMEANS TRTP / CL PDIFF DIFF;
RUN;

DATA DIFF1;
    SET DIFF;
    COMPARISON= TRTP||' VS '||LEFT(_TRTP);
RUN;

PROC SORT DATA=DIFF1; BY COMPARISON_IMPUTATION_; RUN;

PROC MIANALYZE DATA=MIXEDLSMEANS1;
ODS OUTPUT PARAMETERESTIMATES=GROUP_OUTPUT;
BY TRTP;
MODELEFFECTS ESTIMATE;
```

```
STDERR STDERR;
RUN;

PROC MIANALYZE DATA=DIFF1;
    ODS OUTPUT PARAMETERESTIMATES=GROUP_OUTPUT;
    BY comparison;
    MODELEFFECTS ESTIMATE;
    STDERR STDERR;
RUN;
```

Note: The input dataset ALL includes all 20 imputed datasets. TRTP denotes the treatment group and STRATA denotes the stratification factor used in analysis. CHG denotes the change from baseline value and BASELINE denotes the baseline value.

Appendix E. Exposure Adjusted AE Rate Difference and Normal Approximation Based 95% Confidence Interval[6]

Assume the occurrence of TEAE of special interest follows a Possion distribution and let λ denote the rate of occurrence of TEAE under the total exposure time for a treatment group. Let n_1 and n_2 be the number of AEs reported in an upadacitinib dose group and the combined placebo group, respectively. Let T_1 and T_2 be the total number of days exposed to study drug summed across all treated subjects in an upadacitinib dose group and the combined placebo group. Under the assumption that n1 and n2 follow independent Possion distribution with parameters $\lambda_1 T_1$ and $\lambda_2 T_2$, the $\hat{\lambda}_1 = n_1/T_1$ and $\hat{\lambda}_2 = = n_2/T_2$. So the exposure adjusted TEAE rate difference can be estimated by

$$\theta = \hat{\lambda}_1 - \hat{\lambda}_2$$

Using normal approximation, the 95% confidence interval can be calculated by

$$\hat{\lambda}_1 - \hat{\lambda}_2 \pm Z_{\alpha/2}\hat{\sigma}$$

Where $\hat{\sigma} = \sqrt{n_1/T_1^2 + n_2/T_2^2}$

Appendix F. Statistical Analysis to Account for the Impact of COVID-19 Pandemic

This section presents the changes in statistical analysis to handle the missing data due to COVID-19 and the additional data collected related to COVID-19. The Week 24 primary database lock was completed in Jan 2020. The analyses specified in this section only applies to the long-term database lock.

Patient Disposition

Reasons for premature discontinuation of study drug/study will be updated to add two additional reasons related to COVID-19.

- COVID-19 Infection
- COVID-19 Logistical restrictions

The number and percentage of subjects who prematurely discontinued from study drug/study due to COVID-19 infection or COVID-19 logistical restrictions will be summarized by randomized treatment group.

Summary of COVID-19 Impacted Visits

Types of missing visits related to COVID-19 will be collected for the protocol pre-specified visits. For each visit, the number and percentage of subjects impacted by COVID-19 will be summarized by the types of visit for each randomized treatment group as well as overall:

- In person, partial assessments done
- Virtual visit
- Missed visit

Efficacy Analyses

With respect to COVID-19 pandemic, it would be reasonable to assume that missed visits and missing data due to self-quarantine or local government restrictions on travel or limitations on healthcare resources will impact subjects across treatment arms in a similar fashion and adopt the Missing at Random (MAR) assumption for missing data handling.

As described in Section 9.3, longitudinal analysis models MMRM and GLMM, as well as NRI analysis for binary variables will be used for long term efficacy analysis for the treatment sequences, and will be maintained for long term efficacy analysis in the presence of missing data due to COVID-19.

Additionally, treatment comparisons between each upadacitinib dose versus adalimumab will be made for non-radiographic efficacy endpoints for the originally randomized upadacitinib groups and adalimumab group.

For radiographic related endpoints, the analysis will be based on linear extrapolation imputation (Week 24 and Week 56 analysis) and AO data. MMRM will be used for continuous radiographic endpoints for Week 104 and future analysis.

Safety Analyses

COVID-19 related AEs occurring on or after February 1, 2020 and COVID-19 related deaths occurring on or after February 1, 2020 will be collected.

Listings of subjects with AEs related to COVID-19 infection will be provided.