1.0 Title Page

Statistical Analysis Plan

Study M16-098

A Multicenter, Randomized, Double-Blind, Placebo-Controlled Study Evaluating the Safety and Efficacy of Upadacitinib in Subjects with Active Ankylosing Spondylitis

Date: 20 August 2020

Version 3.0

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 M16-098. It provides details to further elaborate on statistical methods as outlined in the protocol. Pharmacokinetic and [redacted] analyses will be performed separately, and the corresponding analysis plan is documented separately.

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

4.0 Study Objectives, Design and Procedures

4.1 Study Objectives

Period 1

1. To evaluate the efficacy of upadacitinib compared with placebo on reduction of signs and symptoms as measured by proportion of subjects who achieve an Assessment of SpondyloArthritis international Society (ASAS) 40 response at Week 14 in subjects with active ankylosing spondylitis (AS) who have had an inadequate response to at least two nonsteroidal anti-inflammatory drugs (NSAIDs) or intolerance to or a contraindication for NSAIDs, and who are biologic disease-modifying antirheumatic drug (bDMARD)-naïve.
2. To assess the safety and tolerability of upadacitinib in subjects with active AS who have had an inadequate response to at least two NSAIDs or intolerance to or a contraindication for NSAIDs, and who are bDMARD-naïve.

Period 2

To evaluate the safety, tolerability, and efficacy of upadacitinib through up to 2 years of treatment in subjects who have completed Period 1 under global protocol V2. After Week 104, subjects in France, Germany, Belgium, Finland, and Netherlands only will be

evaluated for safety and tolerability of upadacitinib every 12 weeks until the end of the study as defined in the country specific protocols V2.01, V2.02 and V2.03.

4.2 Overall Study Design and Plan

This is a Phase 2/3, multicenter study that includes two periods. Period 1 is the 14-week randomized, double-blind, parallel-group, placebo-controlled period designed to compare the safety and efficacy of Upadacitinib 15 mg QD versus placebo for the treatment of signs and symptoms of subjects with active AS who have had an inadequate response to at least two NSAIDs over an at least 4-week period in total at maximum recommended or tolerated doses or intolerance to or a contraindication for NSAIDs as defined by the Investigator, and who are bDMARDs-naïve. Period 2 is an open label long-term extension to evaluate the long-term safety, tolerability, and efficacy of Upadacitinib 15 mg QD in subjects with AS who have completed Period 1.

The study is designed to enroll approximately 170 subjects at approximately 110 study centers worldwide to meet scientific and regulatory objectives without enrolling an undue number of subjects in alignment with ethical considerations.

The study duration includes a 35-day screening period; a 14-week randomized, double-blind, placebo-controlled treatment period (Period 1); a 90-week open label extension period (Period 2); and a 30-day follow-up visit. After Week 104, subjects in France, Germany, Belgium, Finland, and Netherlands will be evaluated for safety and tolerability of upadacitinib every 12 weeks until the end of the study.

X-rays of the pelvis will be performed within the 35-day screening period to evaluate the SI joints to confirm the fulfillment of the modified New York Criteria for AS. X-rays of the spine will also be performed within the 35-day screening period to assess for total spinal ankylosis; subjects with total spinal ankylosis are not eligible for this study. The x-rays of the spine and pelvis will not be required during the Screening Period if the subject had a previous anteroposterior (AP) pelvis x-ray and lateral spine x-rays within

90 days of the Screening Period, provided that the x-rays are confirmed to be adequate for the required evaluations and are deemed acceptable by the central imaging vendor.

Subjects who meet eligibility criteria will be randomized in a 1:1 ratio to one of two treatment groups:

- Group 1: Upadacitinib 15 mg QD, N = 85 (Day 1 to Week 14) → Upadacitinib 15 mg QD (Week 14 and thereafter)
- Group 2: Placebo, N = 85 (Day 1 to Week 14) → Upadacitinib 15 mg QD (Week 14 and thereafter)

Randomization is stratified by screening hsCRP (≤ ULN vs. > ULN) and geographic region (United States [US]/Canada, Japan, Rest of the World [RoW]).

Starting at Week 16, subjects who do not achieve at least an ASAS 20 response at two consecutive visits will have the option to add or modify doses of NSAIDs, acetaminophen/paracetamol, low potency opioid medications (tramadol or combination of acetaminophen and codeine or hydrocodone), and/or modify dose of MTX or SSZ at Week 20 or thereafter. ASAS 20 calculation for rescue therapy no longer applies post Week 104.

Starting at Week 24, subjects who still do not achieve at least an ASAS 20 response at two consecutive visits will be discontinued from study drug treatment. ASAS 20 calculation for discontinuation criteria no longer applies post Week 104.

Subjects who complete the Week 14 visit (end of Period 1) will enter the open-label long-term extension portion of the study, Period 2. Subjects who are assigned to Upadacitinib in Period 1 will continue to receive Upadacitinib 15 mg QD in an open-label manner. Subjects who were randomized to placebo at Baseline will also receive open label upadacitinib 15 mg QD at Week 14.

The primary analysis will be conducted after all subjects have completed Week 14 or have prematurely discontinued prior to Week 14.

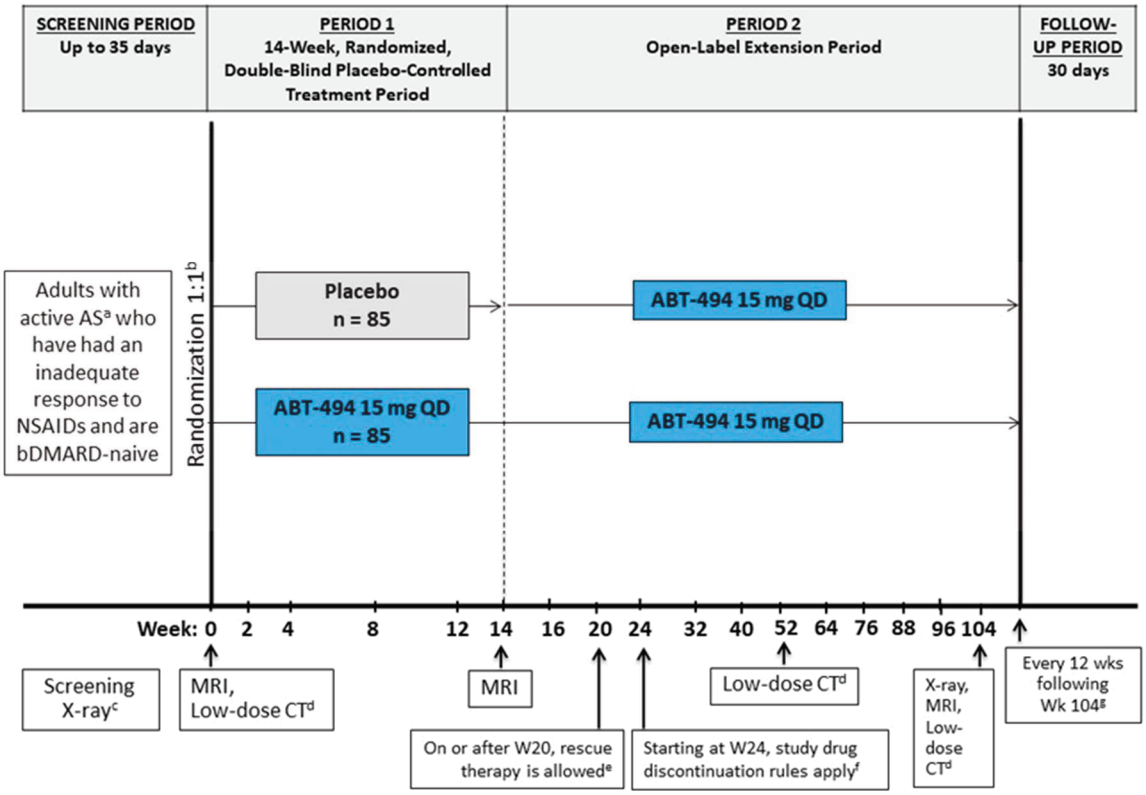
Optional samples may be collected for exploratory research at designated time points throughout the study.

Subjects will have an x-ray of the spine at Week 104. All subjects who meet eligibility criteria will have an MRI evaluation of the SI joints, as well as the cervical, thoracic, and lumbar regions of the spine, prior to or at the Baseline Visit, at Week 14, and Week 104.

Subjects at select sites who consent to participate in the low-dose computer tomography (CT) scan sub-study and meet eligibility criteria will have low-dose CT scan evaluation of the whole spine (cervical, thoracic, and lumbar spine) prior to or at the Baseline Visit, Week 52, and Week 104.

Study design schematic is shown in Figure 1.

Figure 1. Study Design



AS = ankylosing spondylitis; ASAS = Assessment of SpondyloArthritis international Society; BASDAI = Bath Ankylosing Spondylitis Disease Activity Index; bDMARD = biologic disease-modifying antirheumatic drugs; CT = computer tomography; hsCRP = high sensitivity C-reactive protein; MRI = magnetic resonance imaging; NRS = numeric rating scale; NSAIDs = nonsteroidal anti-inflammatory drugs; QD = once daily; RoW = Rest of the World; SSZ = Sulfasalazine; ULN = upper limit of normal; W = week

a. Clinical diagnosis of AS and meeting the modified New York Criteria for AS. Subject must have baseline disease activity as defined by having BASDAI score ≥ 4 and Patient's Assessment of Total Back Pain score ≥ 4 based on a 0 – 10 NRS at the Screening and Baseline Visit.

b. Stratified by geographic region (US/Canada, Japan, RoW) and hsCRP (≤ ULN vs. > ULN).

c. The x-rays of the spine and pelvis will not be required during the Screening Period if the subject had a previous anteroposterior pelvis x-ray and lateral spine x-rays within 90 days of the Screening Period, provided that the x-rays are confirmed to be adequate for the required evaluations and are deemed acceptable by the central imaging vendor.

d. For subjects at select sites who consent to participation in the low-dose spine CT scan sub-study.

e. Starting at Week 16, subjects who do not achieve at least an ASAS 20 response at two consecutive visits will have the option to add or modify doses of NSAIDs, acetaminophen/paracetamol, low potency opioid medications (tramadol or combination of acetaminophen and codeine or hydrocodone), and/or modify dose of MTX or SSZ at Week 20 or thereafter. ASAS 20 calculation for rescue therapy no longer applies post Week 104.

Figure 1. Study Design (Continued)

f. Starting at Week 24, subjects who still do not achieve at least an ASAS 20 response at two consecutive visits will be discontinued from study drug treatment. ASAS 20 calculation for discontinuation criteria no longer applies post Week 104.

g. After Week 104, subjects in France, Germany, Belgium, Finland, and Netherlands will be evaluated for safety and tolerability of upadacitinib every 12 weeks until the end of the study.

4.3 Sample Size

The planned total sample size of 170 can provide at least 90% power to detect a 26% difference in ASAS 40 response rates at Week 14 (assuming a placebo ASAS 40 response rate of 20%), at two-sided $\alpha = 0.05$ and accounting for 10% dropout rate.

4.4 Week 14 Analysis and Database Lock

An unblinded analysis will be conducted after all subjects have completed Period 1 (Week 14) or have prematurely discontinued prior to Week 14.

4.5 Long-term Analysis and Database Locks

A Week 104 analysis will be conducted after all subjects have completed Week 104 or have prematurely discontinued prior to Week 104. This will be the final efficacy analysis for the entire study. Safety analysis will include data collected up to Week 104 or premature discontinuation for every subject.

At the final database lock when the study completes, updated long-term safety analysis will be conducted.

4.6 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 recommendations 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.

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 up to Week 14 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.

Major protocol deviations (ICH deviations and other clinically significant non-ICH deviations) will be identified prior to 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 are assigned to a treatment group based on the "as treated" treatment group, regardless of the treatment randomized. The "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 and Table 3 (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 ASAS Components, BASDAI, BASFI and ASDAS components) and Safety Analysis of Vital Signs for Period 1

Protocol Specified Visit Week	Lower Bound	Target Day	Upper Bound
Baseline	–99	1[a]	1
2	2	15	22
4	23	29	43
8	44	57	71
12	72	85	92
14	93	99	First dose date of Period 2

a. Day of first dose of study drug.

Table 2. Analysis Windows for Safety Analysis of lab values for Period 1

Protocol Specified Visit Week	Lower Bound	Target Day	Upper Bound
Baseline	–99	1[a]	1
4	2	29	43
8	44	57	71
12	72	85	92
14	93	99	First dose date of Period 2

a. Day of first dose of study drug.

Table 3. Analysis Windows for Efficacy Analysis for Period 1 (for ASAS HI)

Protocol Specified Visit Week	Lower Bound	Target Day	Upper Bound
Baseline	–99	1[a]	1
2	2	15	22
4	23	29	43
8	44	57	78
14	79	99	First dose date of Period 2

a. Day of first dose of study drug.

Table 4. Analysis Windows for Efficacy Analysis for Period 1 (for AS QoL and MASES)

Protocol Specified Visit Week	Lower Bound	Target Day	Upper Bound
Baseline	–99	1[a]	1
4	2	29	43
8	44	57	78
14	79	99	First dose date of Period 2

a. Day of first dose of study drug.

Table 5. Analysis Windows for Efficacy Analysis for Period 1 (for FACIT-F, ISI, WPAI, BASMI, TJC/SJC and Dactylitis)

Protocol Specified Visit Week	Lower Bound	Target Day	Upper Bound
Baseline	–99	1[a]	1
14	2	99	First dose date of Period 2

a. Day of first dose of study drug.

Table 6. Analysis Windows for Efficacy Analysis for Period 1 (for MRI joints and spine SPARCC)

Protocol Specified Visit Week	Lower Bound	Target Day	Upper Bound
Baseline	–99	1[a]	3
14	4	99	First dose date of Period 2

a. Day of first dose of study drug.

If first dose date of Period 2 is missing, the Upper Bound of Week 14 will be set at Day 106, which is half way of Week 14 and next visit Week 16.

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 (< 40, [40, 65), ≥ 65 years)
- Race (White, Non-White: Black or African American, American Indian or Alaska Native, Native Hawaiian or Other Pacific Islander, Asian, Other)
- Geographic Region (North America, South/Central America, Western Europe, Eastern Europe, Asia, Other)
- Ethnicity (Hispanic or Latino, Not Hispanic or Latino)
- Weight (kg)
- Weight Categories (< 60 kg, ≥ 60 kg)
- Height (cm)
- Body Mass Index (BMI) (kg/m^2)
- Body Mass Index (BMI) Category (kg/m^2) (BMI < 25, BMI ≥ 25)

AS Medical History and Characteristics at Baseline

- Duration (years) of AS symptoms
- Duration (years) since AS diagnosis
- Duration of AS diagnosis categories (< 5 years or ≥ 5 years)
- HLA-B27 (positive, negative)
- Usage and number of different prior NSAIDS (discontinued prior to baseline)
- Usage and number of different baseline NSAIDs (started prior to baseline and ongoing at baseline)
- Number and percentage of previous disorders at baseline (Psoriasis, Inflammatory Bowel Disease, Uveitis)

ASAS and/or ASDAS Components at Baseline

- Patient's assessment of total back pain (BASDAI Question 2 NRS score 0 - 10)
- Patient global assessment of disease activity (NRS score 0 - 10)
- Peripheral pain/swelling (BASDAI Question 3 NRS score 0 - 10)
- Duration of morning stiffness (BASDAI Question 6 NRS score 0 - 10)
- High sensitivity C-reactive protein (hsCRP) (mg/L)
- Patient's assessment of total back pain (NRS score 0 - 10)

- Function – Represented by the BASFI (NRS score 0 - 10)
- Inflammation (mean of items 5 and 6 of BASDAI NRS scores 0 - 10)
- Lateral lumbar flexion from BASMI
- Erythrocyte sedimentation rate (ESR) (mm/hr)

Other Baseline AS Disease Characteristics

- Continuous ASDAS (CRP)
- ASDAS(CRP) categories (ASDAS(CRP) > 3.5 vs ≤ 3.5)
- MRI SPARCC score (Spine)
- MRI SPARCC score (SI joints)
- BASMI
- Continuous MASES
- Presence of Enthesitis (MASES > 0)
- Tender Joint Count (TJC68)
- Swollen Joint Count (SJC66)
- Presence of dactylitis (yes/no)
- Total dactylitis count out of subjects with presence of dactylitis
- mSASSS

Patient Reported Outcomes at Baseline

- BASDAI
- FACIT-F
- ISI total score and score categories
- WPAI
- AS QoL
- ASAS HI
- Patient's Assessment of Nocturnal Back Pain (NRS score 0 - 10)
- Physician's Global Assessment of Disease Activity (NRS score 0 - 10)
- Patient's Global Assessment of Pain (NRS score 0 - 10)

Clinical Tests at Screening

- Chest x-ray
- 12-Lead ECG
- AP Pelvis X-Ray
- Tuberculosis test result (PPD positive or QuantiFERON positive)
- Anterior Uveitis

Immunization History

- BCG immunization
- Herpes Zoster immunization
- Hepatitis B immunization

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 body systems and conditions/diagnoses as captured on the CRF. The body systems will be presented in alphabetical order and the conditions/diagnoses will be presented in alphabetical order within each body system. The number and percentage of subjects with a particular condition/diagnosis will be summarized for each randomized treatment group as well as overall. Subjects reporting more than one condition/diagnosis within a body system will be counted only once for that body system. 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. This includes medications with a start date between first study drug administration and last study drug administration + 1 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 concomitant NSAIDs will also be reported using equivalent NSAID score and accumulative NSAID intake score (the derivation is detailed in Appendix A).

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

Subjects may have more than one reason for discontinuing; both the primary reason and all reasons will be summarized.

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

Period 2

Period 2 patient disposition and reason for discontinuation will be summarized for overall Upadacitinib 15 mg QD treatment in Period 2.

Among the subjects who entered Period 2 participation (regardless of whether subject prematurely discontinued study drug in Period 1), the number and percentage of subjects 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 completed and prematurely discontinued study drug in Period 2 will be summarized.

For subjects who prematurely discontinued study drug or study participation, the reason for discontinuation will be summarized by the following categories (as collected in CRF):

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

Subjects may have more than one reason for discontinuing; both the primary reason and all reasons will be summarized.

8.0 Study Drug Exposure and Compliance

8.1 Study Drug Exposure

The duration of exposure to study drug will be summarized for the safety analysis set by the randomized treatment groups (placebo group vs Upadacitinib 15 mg QD) in Period 1. For long term, the duration of exposure to study drug will be summarized for the safety analysis set only for the Any Upadacitinib 15 mg QD group, which includes Upadacitinib

15 mg QD exposure from subjects starting on Upadacitinib 15 mg QD and subjects switching from placebo to Upadacitinib 15 mg QD.

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

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

8.2 Compliance

Study drug compliance will be summarized for each treatment group for double blind period up to Week 14. The 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 divided by the number of days that the subject was in treatment.

9.0 Efficacy Analysis

9.1 General Considerations

There are two sets of planned efficacy analysis: efficacy analysis for Period 1 and long-term efficacy analysis. All efficacy analyses will be carried out using the FAS population.

9.1.1 Efficacy Analysis at Different Phases of the Study

Efficacy Analysis for Period 1

Standard efficacy analysis by randomized treatment groups (Upadacitinib 15 mg QD and the placebo group) will be performed on efficacy data for Period 1. No protocol-defined treatment switching will occur in Period 1. 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. Upadacitinib 15 mg QD → Upadacitinib 15 mg QD

There will be no statistical testing; only descriptive statistics and confidence intervals will be provided.

9.1.2 Handling for Missing Data and Intercurrent Event

Non-Responder Imputation (NRI) Approach

The NRI approach will handle data for binary variables 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 binary variables at a specific visit will be treated as non-responder for that visit.

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

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, 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 variables (refer to Section 9.2.2 and Section 9.2.3 where the corresponding supplementary estimand for the primary and ranked secondary endpoints is described respectively).

Mixed Effect Model Repeat Measurement (MMRM)

The repeated measure analysis will be conducted using mixed model 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 and treatment-by-visit interaction, main stratification factors at randomization 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 and using the method of restricted maximum likelihood (REML). MMRM will be used for the primary estimand of continuous variables (refer to Section 9.2.3).

Longitudinal Analysis for Long-Term Efficacy: MMRM and Generalized Linear Mixed Model (GLMM)

The repeated measure analysis will be conducted using mixed model including AO measurements at all visits. MMRM will be used for continuous endpoints and GLMM will be used for binary endpoints. The mixed models will include the categorical fixed

effects of treatment, visit and treatment-by-visit interaction, and stratification factor hsCRP level (≤ ULN vs. > ULN). For the MMRM analysis of change from baseline in continuous endpoints, the baseline measurement will be included as a continuous fixed covariate. The categorical fixed effect of subject's discontinuation status and other baseline covariates may also be included in the model as appropriate. Unstructured, Toeplitz, compound symmetry, or other covariance structures may be considered.

9.2 Efficacy Analysis for Period 1

9.2.1 Primary Efficacy Analysis

The primary efficacy endpoint for regulatory purposes is ASAS 40 response at Week 14. The primary estimand is the difference in the proportion of AS patients who achieved ASAS 40 response at Week 14 and did not discontinue study drug by Week 14, comparing those who are randomized to the Upadacitinib 15 MG QD 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 and the placebo group). 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 Upadacitinib and the placebo group using the Cochran-Mantel-Haenszel (CMH) test adjusting for stratification factor of hsCRP level (≤ ULN vs. > ULN). Point estimate, 95% CI using normal approximation and p-value for the treatment comparison will be presented. The nominal p-value 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. For the primary estimand, NRI data handling as defined in Section 9.1.2 will be used. To facilitate the interpretation of the estimand, the number and percentage of non-responders for ASAS40 will be summarized into three categories:

1. Subjects who discontinue study drug by Week 14

2. Subjects who didn't discontinue study drug but missing Week 14 ASAS 40 measurements
3. Subjects with ASAS 40 measurements observed and on study drug at Week 14 but didn't meet ASAS 40 response criteria

9.2.2 Supplementary Analysis of Primary Efficacy Variable

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 AS patients who achieved ASAS40 response at Week 14, regardless of whether the subject had discontinued study drug by Week 14, comparing upadacitinib 15 MG QD vs Placebo for those who are randomized, received study drug and have the efficacy measurement at Week 14 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 conducted for the primary endpoint. Details of the analysis are outlined in Appendix E.

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 key multiplicity adjusted secondary efficacy endpoints at Week 14 are:

1. Change from Baseline in Ankylosing Spondylitis Disease Activity Score (ASDAS(CRP));
2. Change from Baseline in MRI Spondyloarthritis Research Consortium of Canada (SPARCC) score (Spine);

3. Proportion of subjects with BASDAI 50 response (defined as 50% improvement in the Bath AS Disease Activity Index);
4. Change from Baseline in AS QoL;
5. Proportion of subjects with ASAS partial remission (PR) (defined as an absolute score of ≤ 2 units for each of the four domains identified in ASAS 40);
6. Change from Baseline in BASFI;
7. Change from Baseline in $BASMI_{lin}$;
8. Change from Baseline in MASES (for subjects with baseline enthesitis);
9. Change from Baseline in WPAI (the overall work impairment due to SpA);
10. Change from Baseline in ASAS HI.

Additional key secondary endpoints are:

- ASAS 20 response at Week 14.
- Change from Baseline in MRI SPARCC score (SI joints) at Week 14.

For binary endpoints, frequencies and percentages will be reported for each randomized treatment group. 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. NRI data handling will be used to analyze the primary estimand.

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. To explore various missing data assumptions including missing not at random (MNAR), tipping point analysis will also be conducted for ASAS 20 at Week 14 and details of the analysis are outlined in Appendix E.

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

upadacitinib and the placebo group using the Cochran-Mantel-Haenszel. The CMH test adjusts for the main stratification factor of hsCRP level (≤ ULN vs. > ULN). Point estimate, 95% CI using normal approximation and the p-value for the treatment difference will be presented.

For continuous key secondary efficacy endpoints, the estimand is the difference in mean change from baseline at Week 14 under the assumption that patients with missing data including those due to premature discontinuation of study drug can have their measurements at Week 14 predicted by their observed data and the observed data for other patients for their respective assessments during follow-up. The comparison is Upadacitinib 15 MG QD vs placebo for patients randomized and treated with at least one dose of study drug.

For the primary estimand of the continuous key 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 the main stratification factor of hsCRP level (≤ ULN vs. > ULN). The LS mean and 95% CI will be reported for each randomized treatment group; the LS mean treatment difference and associated 95% CI and p-value will be reported comparing Upadacitinib with the placebo group.

The supplementary analysis for secondary continuous variables will be conducted on AO cases using the analysis of covariance (ANCOVA) model with treatment and the stratification factor of hsCRP (≤ ULN vs. > ULN) as the fixed factor and the corresponding baseline value as the covariate. The corresponding estimand is the difference in the mean change from baseline in the efficacy endpoints at Week 14 regardless of premature discontinuation of study drug. The comparison will be the Upadacitinib 15 MG QD vs placebo for patients randomized and treated with at least one dose of study drug and have the efficacy measurement available at Week 14. To explore various missing data assumptions including missing not at random (MNAR), the tipping point analysis will also be conducted using multiple imputation (MI) as additional supplementary analysis. Details of the tipping point analysis are outlined in Appendix E.

9.2.4 Exploratory Efficacy Analyses

Additional endpoints are the following measurements assessed in subjects treated with upadacitinib versus placebo at scheduled time points other than those specified for the primary and key secondary variables:

- Proportion of subjects with:
  - ASAS 20 response;
  - ASAS 40 response;
  - ASAS PR;
  - ASAS 5/6 (20% improvement from Baseline in five out of the following six domains: BASFI, patient's assessment of total back pain, PtGA, inflammation [mean of items 5 and 6 of the BASDAI] lateral lumbar flexion from $BASMI_{lin}$, and high sensitivity CRP [hsCRP]);
  - Inactive Disease based on ASDAS(CRP) and ASDAS(ESR) (ASDAS score < 1.3);
  - Low Disease based on ASDAS(CRP) and ASDAS(ESR) (ASDAS score < 2.1);
  - Major Improvement based on ASDAS(CRP) and ASDAS(ESR) (a change from Baseline ≤ –2.0);
  - Clinically Important Improvement based on ASDAS(CRP) and ASDAS(ESR) (a change from Baseline ≤ –1.1);
  - Resolution of dactylitis (for subjects with baseline presence of dactylitis);
- Change from Baseline in:
  - ASAS HI;
  - ASDAS(CRP) and ASDAS(ESR) respectively;
  - AsQoL;
  - BASDAI;
  - BASFI;
  - $BASMI_{lin}$;
  - CRP;

- Total dactylitis count (for subjects with baseline presence of dactylitis);
- FACIT-F;
- ISI;
- MASES (for subjects with baseline MASES > 0);
- Modified Stoke Ankylosing Spondylitis Spine Score (mSASSS) score with conventional radiograph;
- MRI SPARCC score of SI joints;
- MRI SPARCC score of Spine;
- Patient's Assessment of Total Back Pain NRS score 0 - 10;
- Patient's Assessment of Nocturnal Back Pain NRS score 0 - 10;
- Patient's Global Assessment of Pain NRS score 0 - 10;
- Physician's Global Assessment of Disease Activity NRS score 0 - 10;
- Inflammation (mean of items 5 and 6 of BASDAI NRS scores 0 - 10);
- Patient's assessment of total back pain (BASDAI Question 2 NRS score 0 - 10);
- Peripheral pain/swelling (BASDAI Question 3 NRS score 0 - 10);
- Duration of morning stiffness (BASDAI Question 6 NRS score 0 - 10);
- Patient's Global Assessment of Disease Activity NRS score 0 - 10;
- TJC68 and SJC66;
- WPAI (all 4 dimension scores);

- Categories in ISI

For binary endpoints, point estimate and 95% CI using normal approximation will be provided for the response rate for each randomized treatment group. Point estimate, 95% CI and p-value will be provided for the treatment comparison between Upadacitinib and the placebo group, where the 95% CI will be based on normal approximation and the p-value will be based on the CMH test adjusting for stratification factor of hsCRP level. Only the nominal p-value will be provided. NRI data handling will be used.

For continuous endpoints, the LS mean and 95% CI will be reported for each randomized treatment group. The LS mean treatment difference and associated 95% CI and p-values between Upadacitinib and the placebo group will be provided using MMRM model with fixed effects of treatment, visit and treatment-by-visit interaction, hsCRP level (≤ ULN vs. > ULN) and baseline value as covariate. Only the nominal p-value will be provided.

For the ASAS components (the 4 components are Patient's Global Assessment of Disease Activity NRS score; Patient's Assessment of Total Back Pain NRS score; BASFI; Inflammation (mean of items 5 and 6 of BASDAI NRS scores 0 - 10)), supplementary analysis will be conducted using AO data handling. Analysis of covariance (ANCOVA) model will be employed with treatment and the stratification factor of hsCRP level (≤ ULN vs. > ULN) as the fixed factor and the corresponding baseline values as the covariate.

For categorical (more than 2 categories) endpoints, number and percentage for each category will be provided for each randomized treatment group. P-value will be provided for the treatment comparison between Upadacitinib and the placebo group using the extended Cochran-Mantel-Haenszel test adjusting for stratification factor of hsCRP level (≤ ULN vs. > ULN). Only the nominal p-value will be provided. No missing data imputation will be used.

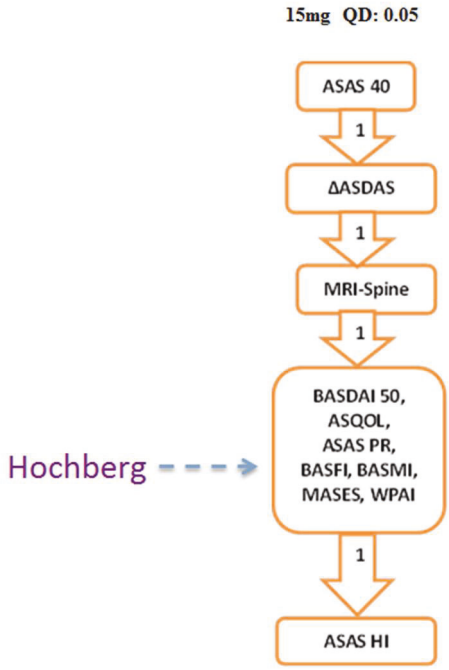
Evaluation of structural changes in the spine will be explored in subjects who are participating in the low-dose CT scan sub-study using a scoring system for syndesmophytes. Analysis for this sub-study is described in Appendix G.

9.2.5 Handling of Multiplicity

In order to preserve Type I error, a step-down approach will be used to test the primary and ranked key secondary endpoints where statistical significance can be claimed for a lower ranked endpoint only if the previous endpoint in the sequence meets the requirements of significance. Specifically, the testing will utilize the endpoint sequence of primary endpoint followed by ranked key secondary endpoints in the order as specified

in following Figure 2, using α of 0.05. The group of multiple endpoints (including proportion of subjects with BASDAI 50 response, proportion of subjects with ASAS partial remission, changes from baseline in AS QoL, BASFI, $BASMI_{lin}$, MASES, and WPAI) will be tested using the Hochberg procedure, conditional on significance of higher-ranked endpoints.

Figure 2. Graphical Multiple Testing Procedure



9.2.6 Efficacy Subgroup Analysis

The primary efficacy endpoint will be examined in the subgroups listed in Table 7 below. Treatment difference between Upadacitinib and the placebo group will be presented with point estimate and 95% confidence interval using normal approximation. No p-value will be provided for subgroup analysis. If any of the resulting subgroups has fewer than 20%

of the planned study size (i.e., < 34 subjects), the subgroup analyses for that variable will not be presented. A forest plot will be provided for the subgroup analyses.

Table 7. Subgroups for Efficacy Analysis

Subgroup Factor	Categories
Age	< 40, [40, 65), ≥ 65
Sex	Male or Female
BMI	< 25, ≥ 25
Race	White vs non-White
Geographic Region	North America, Europe, Other
hsCRP level at screening	≤ ULN vs > ULN

9.3 Long-Term Efficacy Analysis

Assessments to evaluate long-term efficacy will be analyzed for all efficacy measurements at scheduled visits.

Descriptive statistics 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, 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 primary and ranked secondary endpoints. These efficacy analyses will be based on As Observed (AO) analysis. In addition, longitudinal analysis will be performed using MMRM or GLMM as described in Section 9.1.2 for all endpoints except for radiographic endpoints. Point estimates and 95% CI from the model will be provided for each treatment group sequence.

9.4 Efficacy Variables Definitions and Conventions

9.4.1 Ankylosing Spondylitis Disease Activity Score (ASDAS)

Parameters used for the calculation of ASDAS:

1. Patient's assessment of total back pain (BASDAI Question 2 NRS score 0 - 10)

2. Patient global assessment of disease activity NRS score 0 - 10
3. Peripheral pain/swelling (BASDAI Question 3 NRS score 0 - 10)
4. Duration of morning stiffness (BASDAI Question 6 NRS score 0 - 10)
5. high-sensitivity C-reactive protein (hs-CRP) in mg/L.

Calculation of ASDAS:

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

Note: When the conventional CRP is below the limit of detection or when the high sensitivity CRP is < 2 mg/L, the constant value of 2 mg/L should be used to calculate ASDAS-CRP

$ASDAS_{ESR}$ = 0.113 × patient global + 0.293 × √ESR + 0.086 × peripheral pain/swelling + 0.069 × duration of morning stiffness + 0.079× total back pain.

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

- ASDAS Inactive Disease: ASDAS < 1.3
- ASDAS Low Disease: 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 ≤ –2.0)
- ASDAS Clinically Important Improvement (a change from baseline ≤ –1.1)

9.4.2 Bath Ankylosing Spondylitis Disease Activity Index (BASDAI)

BASDAI consists of the following 6 questions; Questions 1 through 5 have responses that can range from 0 (none) to 10 (very severe); Question 6 have response range from 0 (0 hr) to 10 (2 or more hrs), and 5 represents 1 hr.

Q1. How would you describe the overall level of fatigue/tiredness you have experienced?

Q2. How would you describe the overall level of AS neck, back or hip pain you have had?

Q3. How would you describe the overall level of pain/swelling in joints, other than neck, back or hips you have had?

Q4. How would you describe the overall level of discomfort you have had from any areas tender to touch or pressure?

Q5. How would you describe the overall level of morning stiffness you have had from the time you wake up?

Q6. How long does your morning stiffness last from the time you wake up?

Scoring of the BASDAI

BASDAI will be reported 0 to 10. The score has a maximum value of 10 and is calculated as follows:

BASDAI Score = 0.2 (Q1 + Q2 + Q3 + Q4 + Q5/2 + Q6/2)

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.

9.4.3 BASDAI50

BASDAI50 is a categorical response based on BASDAI that represents a 50% improvement in BASDAI from baseline.

9.4.4 ASAS20, ASAS40, ASAS 5/6, and ASAS Partial Remission

Parameters used for the ASAS responses:

- Patient's Global Assessment – Represented by the PTGA-disease activity (NRS score 0 - 10)
- Pain – Represented by the patient's assessment of total back pain (NRS score 0 - 10)
- Function – Represented by the BASFI (NRS score 0 - 10)
- Inflammation – Represented by the mean of the 2 morning stiffness-related BASDAI (mean of items 5 and 6 of the BASDAI NRS score 0 – 10)

ASAS20 Response

Improvement of ≥ 20% and absolute improvement of ≥ 1 unit (on a scale of 0 to 10) from Baseline in ≥ 3 of the above 4 domains, with no deterioration in the remaining domain (defined as a worsening of ≥ 20% and a net worsening of ≥ 1 unit).

ASAS40 Response

Improvement of ≥ 40% and absolute improvement of ≥ 2 units (on a scale of 0 to 10) from Baseline in ≥ 3 of the 4 domains above with no deterioration (defined as a net worsening of > 0 units) in the potential remaining domain.

ASAS Partial Remission

Absolute score of ≤ 2 units for each of the 4 domains identified above.

ASAS 5/6 Response

Improvement of ≥ 20% from Baseline in 5 out of the following 6 domains: BASFI, patient's assessment of total back pain, PTGA-disease activity, inflammation (mean of Items 5 and 6 of the BASDAI]), lateral lumbar flexion from BASMI, and hs-CRP.

9.4.5 Bath Ankylosing Spondylitis Functional Index (BASFI)

The Bath Ankylosing Spondylitis Functional Index (BASFI) consists of the following 10 questions, each with a response ranging from 0 (easy) to 10 (impossible).

1. Putting on your socks or tights without help or aids (e.g., sock-aid).
2. Bending forward from the waist to pick up a pen from the floor without an aid.
3. Reaching up to a high shelf without help or aids (e.g., helping hand).
4. Getting up out of an armless dining room chair without using your hands or any other help.
5. Getting up off the floor without help from lying on your back.
6. Standing unsupported for 10 minutes without discomfort.
7. Climbing 12 to 15 steps without using a handrail or walking aid. One foot on each step.
8. Looking over your shoulder without turning your body.

9. Doing physically demanding activities (e.g., physiotherapy, exercises, gardening, or sports).
10. Doing a full day's activities whether at home or at work.

Scoring of BASFI

The BASFI score will be derived based on the average of Questions 1 through 10. If up to 2 items are missing, corresponding scores will be replaced with the mean of the remaining non-missing items. If 3 or more items are missing, BASFI will be considered missing.

9.4.6 High Sensitivity C-Reactive Protein

The high-sensitivity C-reactive protein (hs-CRP) is a laboratory parameter and is an efficacy variable in axial spondyloarthritis. The abnormal and normal values will be determined according to the normal ranges provided by the laboratory for hs-CRP; in this study a normal CRP is defined as ≤ 2.87 mg/L.

9.4.7 Bath Ankylosing Spondylitis Metrology Index (Linear)

The Linear BASMI ($BASMI_{lin}$) composite score will be calculated using the BASMI components.

Scoring of $BASMI_{lin}$

The table below presents the components of $BASMI_{lin}$ and assessment ranges for score.

	Score		
	0	**Between 0 and 10**	**10**
Lateral Lumbar flexion (cm)	A ≥ 21.1	(21.1 – A)/2.1	A ≤ 0.1
Tragus to wall distance (cm)	A ≤ 8	(A – 8)/3	A ≥ 38
Lumbar flexion (modified Schober) (cm)	A ≥ 7.4	(7.4 – A)/0.7	A ≤ 0.4
Intermalleolar distance (cm)	A ≥ 124.5	(124.5 – A)/10	A ≤ 24.5
Cervical rotation (°)	A ≥ 89.3	(89.3 – A)/8.5	A ≤ 4.3

$BASMI_{lin}$ = Assessment measurements for tragus to wall, cervical rotation and lateral lumbar flexion are the means of the left and right measurements

A = assessment measurement

Scores for each assessment range from 0 to 10, and the $BASMI_{lin}$ total score will be the average of the 5 assessment scores. If 1 item is missing, the $BASMI_{lin}$ will be calculated as the mean of remaining 4 items. Hence, the range of the $BASMI_{lin}$ total score should be between 0 and 10. If 2 or more items are missing, then the $BASMI_{lin}$ score will be considered missing.

9.4.8 Maastricht Ankylosing Spondylitis Enthesitis Score (MASES)

The Maastricht Ankylosing Spondylitis Enthesitis Score (MASES) will be measured at the designated study visits listed in the protocol. To assess the presence or absence of enthesitis at 13 different sites, noting the subjects' responses.

Scoring of MASES

The following left and right locations will be graded for presence (1) or absence (0) of enthesitis: 1st Costochondral joint, 7th Costochondral joint, Posterior Superior Iliac Spine, Anterior Superior Iliac Spine, Iliac Crest, Proximal Insertion of Achilles tendon; the 5th Lumbar Spinous process will also be graded for enthesitis yielding a total score ranging 0 – 13. If one or more locations are missing, the score will be calculated using available data. If all locations are missing, then MASES is set to be missing.

9.4.9 Dactylitis

The Dactylitis scores will be the presence of Dactylitis at baseline, the total Dactylitis count (out of subjects with baseline presence of Dactylitis) and the resolution of Dactylitis (out of subjects with baseline presence of Dactylitis).

The presence of Dactylitis at baseline is defined as the following:

At least one affected and tender digit with circumference increase over reference digit $\geq 10\%$.

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 Leeds Dactylitis Index (LDI) is a score based on finger circumference and tenderness, assessed and summed across all dactylitic digits. 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 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 proportion of subjects with resolution of dactylitis is defined as the proportion of subjects with LDI = 0.

9.4.10 Work Productivity and Activity Impairment Questionnaire: Axial Spondyloarthritis, Version 2.0 (WPAI-Axial Spondyloarthritis)

The Work Productivity and Activity Impairment Questionnaire: Axial Spondyloarthritis, V2.0 (WPAI-Axial SpA) was developed to measure the effect of overall health and specific symptoms on productivity at work and outside of work. It consists of 6 questions. A lower WPAI-Axial SpA score indicates an improvement. The

WPAI-Axial SpA is collected at the designated study visits listed in the protocol. The WPAI-Axial SpA coding and scoring methods are described in the following:

A description of the six questions asked in the WPAI-Axial SpA is as follows:

Questions:

Q1. Currently employed? (Yes/No).

Q2. Hours missed from work due to Axial Spondyloarthritis.

Q3. Hours missed due to other reasons.

Q4. Hours actually worked.

Q5. Degree Axial Spondyloarthritis affected productivity while working.

Q6. Degree Axial Spondyloarthritis affected regular activities other than job.

The following 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 above questions.

Scores:

S0. Employment: defined below in missing data handling conventions

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

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

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

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

S3. Percent overall work impairment due to SpA:

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

S4. Percent activity impairment due to SpA:

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

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: Axial SpA 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 or 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.

9.4.11 Joint Evaluation

Anatomical joints are evaluated for swelling and tenderness at the designated study visits listed in the protocol. The 34 anatomical joints in Table 8 are assessed in both the left and right side of the body for tenderness. The joints to be examined for swelling are the same as those examined for tenderness, with the exception of the hip joints.

Table 8. Anatomical Joints Assessed

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 will assess whether a particular joint is "tender or painful," where presence of tenderness is scored as "1" and the absence of tenderness is 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) will be derived as the sum of all "1s" thus collected and no penalty will be considered for the joints not assessed or those which have been replaced. A similar method will be followed for the derivation of total swollen joint count (SJC66). Thus, the range for TJC68 will be 0 to 68 and 0 to 66 for SJC66.

9.4.12 Ankylosing Spondylitis Quality of Life Questionnaire (ASQoL)

Scoring the ASQoL

Each statement on the ASQoL is given a score of "1" or "0." A score of "1" is given where the item is affirmed, indicating adverse QoL. All item scores are summed to give a total score or index. Scores can range from 0 (good QoL) to 18 (poor QoL). Cases with more than three missing responses (i.e., more than 20%) cannot be allocated a total score. For cases with between one and three missing responses, the total score is calculated as

follows: T=18x/18-m where: *T* is the total score, x is the total score for the items affirmed and *m* is the number of missing items.

The 18 statements of the questionnaire are listed below.

Please read each item carefully and tick the one response that applies best to you ***at the moment***	
1. My condition limits the places I can go	Yes □ No □
2. I sometimes feel like crying	Yes □ No □
3. I have difficulty dressing	Yes □ No □
4. I struggle to do jobs around the house	Yes □ No □
5. It's impossible to sleep	Yes □ No □
6. I am unable to join in activities with my friends/family	Yes □ No □
7. I am tired all the time	Yes □ No □
8. I have to keep stopping what I am doing to rest	Yes □ No □
9. I have unbearable pain	Yes □ No □
10. It takes a long time to get going in the morning	Yes □ No □
11. I am unable to do jobs around the house	Yes □ No □
12. I get tired easily	Yes □ No □
13. I often get frustrated	Yes □ No □
14. The pain is always there	Yes □ No □
15. I feel I miss out on a lot	Yes □ No □
16. I find it difficult to wash my hair	Yes □ No □
17. My condition gets me down	Yes □ No □
18. I worry about letting people down	Yes □ No □

9.4.13 ASAS Health Index (HI)

The ASAS HI is a linear composite measure with a dichotomous response option: "I agree" and "I do not agree." Each statement on the ASAS HI is given a score of "1" = "I agree" or "0" = "I do not agree." The total sum of the ASAS HI ranges from 0-17, with a lower score indicating a better health status. Please note that item 7 and 8 are not applicable to all patients. For those patients who ticked the response "not applicable," the sum score is analyzed based on n = 16 or n = 15 respectively. A total score can be analyzed if no more than 20% of the data (i.e., 3 items) are missing. The total score is calculated as follows for respondents with on to a maximum of three missing responses:

Sum.score = x/(17-m)*17, where x is the item summation score and m is the number of missing items and $m \leq 3$. Cases with more than three missing responses ($m > 3$) cannot be allocated a total score and the total score will be set as missing.

The following table lists the 17 items:

Item
1. Pain sometimes disrupts my normal activities.
2. I find it hard to stand for long.
3. I have problems running.
4. I have problems using toilet facilities.
5. I am often exhausted.
6. I am less motivated to do anything that requires physical effort.
7. I have lost interest in sex.
8. I have difficulty operating the pedals in my car.
9. I am finding it hard to make contact with people.
10. I am not able to walk outdoors on flat ground.
11. I find it hard to concentrate.
12. I am restricted in traveling because of my mobility.
13. I often get frustrated.
14. I find it difficult to wash my hair.
15. I have experienced financial changes because of my rheumatic disease.

16. I sleep badly at night.
17. I cannot overcome my difficulties.

9.4.14 Functional Assessment of Chronic Illness Therapy – Fatigue Scale (FACIT-F)

Fatigue is one of the most frequent complaints of the elderly and is strongly associated with loss of independence and decreased physical activity and functional decline. One validated tool to measure fatigue is FACIT Fatigue Scale v4. The FACIT Fatigue Scale is a short, 13-item, easy to administer tool that measures an individual's level of fatigue during their usual daily activities over the past week. Each of the fatigue and impact of fatigue items are measured on a four point Likert scale. The FACIT Fatigue Scale is ranged from 0 to 52 and the higher the score, the better the quality of life.

Score for each item is calculated by either subtracted from 4 or adding 0 depending on whether it is a reversal item or not. FACIT Fatigue Scale is then calculated by adding up all item scores, multiplied by 13 and divided by the number of items answered. It is essentially a prorated subscale if there are missing values for some items. If less than or equal to 50% of the items are answered (e.g., 6 out of 13), the proration is not acceptable and the scale will not be computed.

Below is a list of statements that other people with your illness have said are important. By circling one (1) number per line, please indicate how true each statement has been for you during the past 7 days.

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

9.4.15 Insomnia Severity Index (ISI)

ISI is a 7 item questionnaire that assess, (1) difficulty with sleep onset, (2) difficulty with sleep maintenance (3) problem with early awakening, (4) satisfaction with sleep pattern, (5) interference with daily functioning as a result of sleep problems, (6) noticeability of sleep problem to others, and (7) degree of distress caused by sleep problem. The severity of insomnia items are rated on a Likert type of scale, (0) "None," 1- "Mild," 2- "Moderate," 3- "Severe" and 4- "Very Severe." Other items on satisfaction, noticeability on impairment, worry/distress about sleep and interference are similarly rated on 5-point likert type scale form. See an ISI example below:

For each question, please CIRCLE the number that best describes your answer.

Please rate the CURRENT (i.e., LAST 2 WEEKS) SEVERITY of your insomnia problem(s).

Insomnia Problem	None	Mild	Moderate	Severe	Very Severe
1. Difficulty falling asleep	0	1	2	3	4
2. Difficulty staying asleep	0	1	2	3	4
3. Problem waking up too early	0	1	2	3	4

4. How SATISFIED/DISSATISFIED are you with your CURRENT sleep pattern?

Very Satisfied	Satisfied	Moderately Satisfied	Dissatisfied	Very Dissatisfied
0	1	2	3	4

5. How NOTICEABLE to others do you think your sleep problem is in terms of impairing the quality of your life?

Not at All Noticeable	A Little	Somewhat	Much	Very Much Noticeable
0	1	2	3	4

6. How WORRIED/DISTRESSED are you about your current sleep problem?

Not at All Worried	A Little	Somewhat	Much	Very Much Worried
0	1	2	3	4

7. To what extent do you consider your sleep problem to INTERFERE with your daily functioning (e.g., daytime fatigue, mood, ability to function at work/daily chores, concentration, memory, mood, etc.)?

Not at All Interfering	A Little	Somewhat	Much	Very Much Interfering
0	1	2	3	4

Guidelines for Scoring/Interpretation:

Add the scores for all seven items (questions 1 + 2 + 3 + 4 + 5 + 6 + 7) = _____ your total score. If there is one or two missing items, their value can be replaced with the average

score of the remaining items. If there are more than two items with no responses, it is preferred to consider the total score missing.

Total score categories:

0 – 7 = No clinically significant insomnia

8 – 14 = Subthreshold insomnia

15 – 21 = Clinical insomnia (moderate severity)

22 – 28 = Clinical insomnia (severe)

9.4.16 SpondyloArthritis Research Consortium of Canada (SPARCC) scores

SPARCC scores for spine and sacroiliac (SI) joints are calculated by adding up the dichotomous outcomes from evaluations of the presence, depth and intensity of bone marrow edema lesions of the spine and SI joints, respectively.

SPARCC Assessment of Spine Scoring

In the SPARCC assessment of the MRI of the spine, the entire spine is evaluated for active inflammation (bone marrow edema) using the STIR image sequence.
23 discovertebral units (DVUs) are assessed by each reviewer per subject and timepoint, and the six most severely affected DVUs are selected by each reviewer and used to calculate the MRI Spine SPARCC score. For each of the six DVUs, 3 consecutive sagittal slices are assessed in four quadrants in order to evaluate the extent of inflammation in all three dimensions.

1. Each quadrant is scored for presence of increased signal on STIR
   - 1 = increased signal
   - 0 = normal signal

2. Presence, on each of the sagittal slices, of a lesion exhibiting high signal intensity (comparable to cerebrospinal fluid) in any disco-vertebral unit is given an additional score of 1.
3. Slices that included a lesion demonstrating continuous increased signal of depth ≥ 1 cm extending from the endplate are to be scored as +1 per slice.

The maximum possible score for any individual slice is 6, with a maximum score for all 6 discovertebral units being 108.

SPARCC Assessment of SI Joint Scoring

Scoring is conducted on 6 consecutive slices of the STIR image sequence. All lesions within the iliac bone and within the sacrum up to the sacral foramina are to be scored. The SI joint is divided into 4 quadrants: upper iliac, lower iliac, upper sacral and lower sacral. Each consecutive slice is scored separately for the right and left joint in all four quadrants as follows:

1. Each quadrant is scored for presence of increased signal on STIR
   - 1 = increased signal
   - 0 = normal signal
2. Joints that include a lesion exhibiting intense signal on the STIR sequence are scored as +1 per slice.
3. Joints that included a lesion demonstrating continuous increased signal of depth ≥ 1 cm from the articular surface are be scored as +1 per slice.

The maximum possible score for any individual slice is 12, with a maximum score for all 6 slices being 72.

Adjudication by a third reader

Should adjudication be required for a given case, another reviewer, different from the reviewers who performed primary assessments will make a third, independent assessment. There will be a single dedicated adjudicator performing all adjudication reads.

The 2 closest of the 3 readings (2 primary readers and adjudicator) will determine the final score.

9.4.17 Modified Stoke Ankylosing Spondylitis Spine Score (mSASSS)

The modified Stoke Ankylosing Spondylitis Spine Score (mSASSS) is a scoring method that measure radiographic progression in the spine of patients with ankylosing spondylitis. The mSASSS has a range of 0 to 72, which is derived from scoring the anterior site of the lumbar spine from the lower border of T12 to the upper border of S1 and the anterior site of the cervical spine from the lower border of C2 to the upper border of T1 as either 0 (normal), 1 (erosion, sclerosis, or squaring), 2 (syndesmophyte), 3 (bridging syndesmophyte), or NA vertebral body not evaluable.

Selection of Scores and Adjudication of mSASSS

X-ray of spinal films will be read by two independent readers who are blinded to both treatment and time point. The blinded data read will be performed to compare radiographic progression (from Baseline to the follow-up timepoint, e.g., Year 2).

Adjudication trigger: If the follow-up mSASSS progression scores of read 1 and read 2 differ by [redacted] mSASSS points for a given subject, a third independent reader (adjudicator) will perform an adjudication read and score that subject. The 2 closest of the 3 readings (2 primary reader and adjudicator) will determine the final score.

Handling of missing joints in the mSASSS

If a score at any location (vertebral unit, VU) is missing, the method described below will be used for deriving mSASSS.

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

9.4.18 Other Efficacy Variables

- Patient's Global Assessment of Disease Activity NRS [0 - 10]
  - No activity is indicated by 0 and severe activity by 10.
- Patient's Assessment of Total Back Pain NRS [0 - 10]
  - No pain is indicated by 0 and most severe pain by 10.
- Physician's Global Assessment of Disease Activity NRS [0 - 10]
  - No activity is indicated by 0 and severe activity by 10.
- Patient's Assessment of Nocturnal Back Pain NRS [0 - 10]
  - No pain is indicated by 0 and worst possible pain by 10.
- Patient's Global Assessment of Pain NRS [0 - 10]
  - No pain is indicated by 0 and severe pain by 10.

10.0 Safety Analysis

10.1 General Considerations

Safety analyses will be carried out using the Safety Analysis Set. There are two sets of planned safety analysis: safety analysis for Period 1, and long-term safety analysis.

Safety Analysis for Period 1

Standard safety analysis by the "as treated" treatment groups of Upadacitinib 15 mg QD and placebo group will be performed on safety data in Period 1. No protocol-defined treatment switching will occur prior to these time points.

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 treatment group. All continuous laboratory parameters and vital signs variables at each visit will be summarized by "as treated" treatment group. Frequency tables of subjects meeting criteria for potentially clinically significant vital sign values and for potentially clinically significant laboratory values will be provided by treatment group. Missing safety data will not be imputed.

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, 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. All continuous laboratory parameters and vital signs variables at each visit will be summarized by "as treated" treatment group sequences defined as follows. 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:

1. Placebo → Upadacitinib 15 mg QD
2. Upadacitinib 15 mg QD → Upadacitinib 15 mg QD

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 of the drug after the last dose of study drug.

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. All adverse event tables will be sorted in alphabetical order by SOC and PT and descending percentages for each treatment group.

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 reasonably possibly related to study drug
- TEAEs of special interest

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

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

For TEAEs of special interest, the point estimate and 95% CI (using normal approximation and separate group variance) will be provided for the treatment difference in AE percentage between Upadacitinib group and the placebo group.

10.2.1.2 Adverse Events by System Organ Class and Preferred Term

The number and percentage of subjects experiencing adverse events 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 reasonably possibly related to study drug
- TEAEs leading to discontinuation of study drug
- TEAE leading to death
- Frequent AEs (reported in 2% of subjects or more in any treatment group)

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 that if the subject has another occurrence of the same AE with the most extreme severity – severe. In this case, the subject will be counted under the severe category.

10.2.1.4 TEAEs by Relationship

TEAEs will also be summarized by relationship to Upadacitinib 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 MedDRA PT in decreasing frequency 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 using SOC and MedDRA PT. The AESI categories will be identified by the following search criteria per Standard MedDRA Queries (SMQs)/Company MedDRA Queries (CMQs) in Table 9 below. Adjudicated

cardiovascular events will be summarized and presented by treatment group using the CAC adjudicated categories.

Table 9. 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" Narrow SMQ 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 (Veliparib Product Specific)"

Table 9. AESI for Upadacitinib with SMQs/CMQs/PTs Searches (Continued)

AESI	Type of MedDRA Query	Broad or Narrow Search	SMQ/CMQ Search Criteria
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"
Adjudicated Cardiovascular Events	Output from CAC		
MACE*			
Undetermined/Unknown Cause of Deaths			
Other Cardiovascular events			
Adjudicated Thrombotic Events	Output from CAC		
Venous Thromboembolic Events**			
Other Venous Thrombosis			
Arterial Thromboembolic Events (non-cardiac, non-neurologic)			

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

** Venous thromboembolic events (VTE) include deep vein thrombosis (DVT) and pulmonary embolism (PE).

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

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. According to the study design, randomized placebo subjects switch to Upadacitinib 15 mg

QD when entering Period 2, this analysis will be reported for Any Upadacitinib 15 mg QD subjects. The Any Upadacitinib 15 mg QD treatment group is defined as subjects who receive at least one dose of Upadacitinib 15 mg QD at any time during the study.

This includes AEs occurred under Upadacitinib 15 mg QD exposure from subjects starting on Upadacitinib 15 mg QD and subjects switching from placebo to Upadacitinib 15 mg QD.

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 (days)/denominator])/365.25 • 100 will be presented for each treatment group. The EAER will be the main approach to evaluate AEs in the long-term analysis.

In addition, the exposure-adjusted incidence rate (censored at first event) will be conducted for AESI endpoints as deemed 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 for Any Upadacitinib 15 mg QD 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

The number of TEAEs reported, the total number of years of study drug exposure, and the TEAE rate per 100 patient-years will be presented for Any Upadacitinib 15 mg QD group.

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

For TEAE of special interest, the point estimate and 95% CI (using normal approximation) will be provided for the Any Upadacitinib 15 mg QD group in AE rate per 100 patient-years (refer to Appendix B).

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

For the Any Upadacitinib 15 mg QD treatment group, the TEAE rate per 100 patient-years of exposure as outlined in Section 10.2.2.1 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 reasonably possibly 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 the Any Upadacitinib 15 mg QD treatment group using SOC and MedDRA PT (for adjudicated cardiovascular events, the CAC adjudicated categories will be used).

The AESI categories will be identified per Standard MedDRA Queries (SMQs)/Company MedDRA Queries (CMQs).

The Adverse Events of Special Interest (AESI) rate per 100 patient-years of exposure as outlined in Section 10.2.2.1 will be calculated overall, for each SOC and each PT, for each of the AESI listed in Section 10.2.1.6.

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 10. 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
Uric acid
Sodium
Potassium
Calcium
Inorganic Phosphorus
Creatine Phosphokinase (CPK)
Chloride
Bicarbonate

Table 10. 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
QuantiFERON-TB Gold[a]
ESR

a. For annual follow-up QFT is captured only for those with negative QFT at Screening.

10.3.2 Analysis of Laboratory Data for Period 1

The laboratory data will be summarized by the "as treated" treatment groups (Upadacitinib 15 mg QD and placebo group).

10.3.2.1 Assessment of Mean Change from Baseline and Percentage Change from Baseline in Clinical Laboratory Variables

Analyses of continuous hematology, chemistry, and urinalysis variables which are measured longitudinally will be performed by visits and by treatment group. For each parameter at each visit, the following summary statistics 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 and percent change from baseline between different treatment groups for selected laboratory parameters. Mean difference from placebo and associated 95% CIs will be presented. The analysis applies to the following laboratory parameters of clinical interest: hemoglobin, platelets, lymphocytes, neutrophils, creatinine, creatine phosphokinase (CPK), LDL, HDL, the ratio of LDL to HDL, 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 Low, Normal, and High according to normal ranges. The shift tables will tabulate the number and percentage of subjects with baseline and post-baseline values by these categories.

For each laboratory variable, shift tables will be generated that cross tabulate the subjects' as deemed appropriate by treatment groups:

- Shift from baseline high or normal to minimum post-baseline low; shift from baseline low or normal to maximum post-baseline high.
- Shift from baseline high or normal to final post-baseline low; shift from baseline low or normal to final post-baseline high.

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 3 or higher. The number and percentage of subjects meeting the criteria for potentially clinically significant laboratory values will be summarized by "as treated" treatment group.

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 "as treated" treatment group:

- 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 Mean Change from Baseline and Percentage Change from Baseline in Clinical Laboratory Variables

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

Analyses will be performed for change and percentage change from baseline in hemoglobin, lymphocytes, neutrophils, creatinine, creatine phosphokinase (CPK), LDL, HDL, the ratio of LDL to HDL, and total cholesterol.

10.3.3.2 Assessment of Potentially Clinical Significant Laboratory Values

Long-term laboratory data will also be summarized based on the number and percentage of subjects meeting the criteria for potentially clinical significant laboratory values for the Any Upadacitinib 15 mg QD group as described in Section 10.2.2.

The baseline value will be determined by the last non-missing measurement recorded on or before the date of the first dose of Upadacitinib 15 mg QD. For a subject who started on placebo and switched to upadacitinib 15 mg QD at Week 14, lab values under upadacitinib 15 mg QD exposure would be evaluated against the baseline value defined as above.

A listing of all subjects with any laboratory determination meeting CTCAE criteria of Grade 3 or higher will be provided by grade. 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 for Any Upadacitinib 15 mg QD group as described in Section 10.2.2:

- 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

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, respiratory rate, body temperature, and weight. The criteria for potentially clinically significant vital sign findings are presented in Table 11.

Table 11. 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 ≥ 15 mmHg from Baseline
	High	Value ≥ 105 mmHg and increase ≥ 15 mmHg from Baseline
Pulse	Low	Value ≤ 50 bpm and decrease ≥ 15 bpm from Baseline
	High	Value ≥ 120 bpm and increase ≥ 15 bpm from Baseline
Respiratory Rate	Low	< 10 rpm
	High	> 24 rpm
Body temperature	High	> 39.0°C (102.3°F)
Weight	High	> 7% increase from baseline
	Low	> 7% decrease from baseline

10.4.2 Analysis of Vital Sign for Period 1

Analyses of continuous vital sign variables which are measured longitudinally will be performed by visits and by the treatment groups of Upadacitinib 15 mg QD and 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 "as treated" treatment group.

10.4.3 Analysis of Long-Term Vital Sign

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

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 for the Any Upadacitinib 15 mg QD group as described in Section 10.2.2. Similar baseline definition as described in Section 10.3.3.2 will be applied.

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 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 14 DBL was completed in Feb 2019. The Week 64 database lock was completed in Feb 2020. The analyses specified in this section applies to future database locks.

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 the 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.

A sensitivity analysis will be conducted to handle this using AO data. MMRM will be used for continuous endpoints and GLMM defined in section 9.1.2 will be used for binary endpoints.

Safety Analyses

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

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

12.0 Version History

Table 12. SAP Version History Summary

Version	Date	Summary
1.0	12 Apr 2018	Original version
2.0	20 Dec 2018	Incorporated FDA's comment
3.0	20 August 2020	Incorporating updates 1) the study duration is extended for France, Belgium, Finland, Netherlands and Germany in protocol amendment 2.01, 2.02 and 2.03; 2) Add Week 104 data base lock (DBL) and final lock; 3) in statistical analysis related to COVID-19;4) add appendix G for low dose CT sub-study analysis plan 5) Update ASDAS Low Disease definition for Week 104 DBL 6) Include incident rate for all AESI for future DBL

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. National Cancer Institute. CTCAE v. 4.0. Available from: http://ctep.cancer.gov/protocolDevelopment/electronic_applications/ctc.htm#ctc. Accessed on: August 10, 2017.
3. Dougdos M, Paternotte S, Braun J, et al. ASAS recommendations for collecting, analyzing and reporting NSAID intake in clinical trials/epidemiological studies in axial spondyloarthritis. Ann Rheum Dis. 2011;70(2):249-51.
4. Machado PM, Landewé R, Heijde DV, et al. Ankylosing Spondylitis Disease Activity Score (ASDAS): 2018 update of the nomenclature for disease activity states. Ann Rheum Dis. 2018 Feb 16. pii: annrheumdis-2018-213184. doi: 10.1136/annrheumdis-2018-213184. [Epub ahead of print].
5. Stokes ME, Davis CS, Koch GG. Chapter 4 of Categorical data analysis using SAS. SAS institute. 2012.
6. Liu GF, Wang J, Liu K, et al. Confidence intervals for an exposure adjusted incidence rate difference with applications to clinical trials. Stat Med. 2006;25(8):1275-86.

14.0 Appendix

Appendix A	ASAS-NSAID Intake Score for Axial Spondyloarthritis
Appendix B	Exposure adjusted AE rate and normal approximation based 95% confidence interval
Appendix C	Analysis of Categorical Endpoints
Appendix D	Analysis of Continuous Endpoints using Mixed Effect Model Repeat Measurement (MMRM)
Appendix E	Tipping Point Analysis
Appendix F	Geographic Region
Appendix G	Low Dose CT Sub-study Analysis Plan

Appendix A. ASAS-NSAID Intake Score for Axial Spondyloarthritis

ASAS-NSAID Intake Score for Axial Spondyloarthritis: The amount of NSAID intake could be considered as a clinically relevant outcome measure in ankylosing spondylitis (AS)/axial spondyloarthritis.

To be able to calculate the ASAS-NSAID Intake Score, the following information is required:

(1) the type of NSAID/Cox2 inhibitor: each NSAID has a corresponding weight factor. The weight is determined by the maximum dose of each NSAID to achieve a score of 100. The goal is to analyze/report the data in terms of NSAID equivalent dose in mg/day on a 0 - 100 scale. The 150 mg equivalent diclofenac is set to 100. For instance, 150 mg Diclofenac is equivalent to 1000 mg Naproxen, so the weight for Diclofenac is 100/150, and, the weight for Naproxen is 100/1000. For other NSAIDs/COX2 inhibitors not described in table the weight factor will be determined based on medical review using the local label as a guide.

NSAID	Weights
Diclofenac	100/150
Naproxen	100/1000
Aceclofenac	100/200
Celecoxib	100/400
Etodolac	100/600
Etoricoxib	100/90
Flurbiprofen	100/200
Ibuprofen	100/2400
Indometacin	100/150
Ketoprofen	100/200
Meloxicam	100/15
Nimesulide	100/200
Phenylbutazone	100/400
Piroxicam	100/20
Tenoxicam	100/20

(2) the dose: mg per intake

(3) the frequency: days of intake per week (the times of intake daily)

Frequency Description	Numeric Frequency
QD	1
BID	2
TID	3
QID	4
1 time per week	1/7
2 times per week	2/7
3 times per week	3/7
4 times per week	4/7
5 times per week	5/7
6 times per week	6/7

(4) days of intake during the period of interest: can be calculated by NSAID end date - start date + 1.

(5) days stayed on study in the clinical trial: is defined as the total number of days in the study = last visit date – baseline date + 1.

Then, two scores can be calculated:

1. **Equivalent NSAID score** = weights * dose(mg) * (numeric frequency),

   where weights can be obtained from the weight table (above and Dougados et al. Ann Rheum Dis 2011; 70:249–251), and dose is the number of mg per intake, numeric frequency is the number of intakes per day. The equivalent NSAID score is a scoring system to refer to a scale in which 0 equals no intake, and 100 equals 150 mg diclofenac, or 1000 mg naproxen, or 200 mg aceclofenac, or 400 mg celecoxib etc per daily dose.

2. ASAS-NSAID Intake Score =

(Equivalent NSAID score) * (days of intake during period of interest)/(days stayed on study in the clinical trial)

Appendix B. EAER and Normal Approximation Based 95% Confidence Interval

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 be the number of AEs reported in Any Upadacitinib 15 mg QD group. Let T be the total number of days exposed to study drug summed across all treated subjects in Any upadacitinib 15 mg QD group. Under the assumption that n follow Possion distribution with parameters λT, the $\hat{\lambda}$= n/T.

Using normal approximation, the 95% confidence interval can be calculated by (Liu GF, et al. 2006.):

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

Where $\hat{\sigma} = \sqrt{\frac{n}{T^2}}$

Appendix C. Analysis of Categorical Endpoints

Binary endpoints

The point estimate of the response rate for each treatment group is calculated as $\hat{p} = \frac{k}{N}$ where k denotes the number of responders in the treatment group and N denotes the total number of subjects in the treatment group. The corresponding 95% CI is based on normal approximation, that is $\hat{p} \pm Z_{0.975}\sqrt{\frac{1}{N}\hat{p}(1-\hat{p})}$, where $Z_{0.975}$ denotes the 97.5% quantile of the standard normal distribution. This is implemented by obtaining the Wald asymptotic confidence limits using the SAS procedure PROC FREQ with the BINOMIAL option. The SAS code example is as follows:

```
proc freq data=TESTData1;
  by TRTP;
  tables RESP / binomial(level="1") alpha=0.05;
  weight COUNT/zero;
  output out=CI(keep= L_BIN U_BIN) binomial;
run;
```

Note: The input dataset is cell count data, with variable RESP denoting response status ("1" as responder and "0" as non-responder) and variable COUNT denoting the number of subjects with the corresponding response status.

The treatment difference is calculated as $\hat{p}_1 - \hat{p}_2$, and the corresponding 95% CI for the treatment difference is based on normal approximation, that is $\hat{p}_1 - \hat{p}_2 \pm Z_{0.975}\sqrt{\frac{1}{N_1}\hat{p}_1(1-\hat{p}_1) + \frac{1}{N_2}\hat{p}_2(1-\hat{p}_2)}$. This is implemented by obtaining the Wald asymptotic confidence limits using the SAS procedure PROC FREQ with the RISKDIFF option. The SAS code example is as follows:

```
proc freq data=TESTData1;
  tables TRTP*RESP / riskdiff alpha=0.05;
  weight COUNT/zeros;
  where TRTP in ('PLACEBO','UPA 15MG');
  output out=diff(keep= _RDIF1_ L_RDIF1 U_RDIF1) riskdiff1;
run;
```

Note: The input dataset is cell count data, with variable RESP denoting response status ("1" as responder and "0" as non-responder) and variable COUNT denoting the number of subjects with the corresponding response status. TRTP denotes the treatment group.

The p-value for the treatment comparison is based on the Cochran-Mantel-Haenszel (CMH) test adjusting for stratification factor. This is implemented using the SAS procedure PROC FREQ with the CMH option. The p-value from the "general association statistic" is used. The SAS code example is as follows:

```
proc freq data=TESTData3;
  tables STRATA*TRTP*RESP_VALUE / cmh;
run;
```

Note: The input dataset is subject-level data, with variable RESP_VALUE denoting the subject's response status ("1" as responder and "0" as non-responder). TRTP denotes the treatment group and STRATA denotes the strata for a given subject.

Categorical endpoints with more than 2 categories

The extended Cochran-Mantel-Haenszel test is used to assess the association among a set of q2-by-r tables, where r is the number of levels in the ordinal categorical endpoints. For the case of insomnia severity index (ISI) which has r = 4 levels, the extended CMH test is stratified by hsCRP level that has q = 2 strata. The analysis is to assess the association between treatment and ISI, adjusting for the hsCRP level.

The following PROC FORMAT declaration is to ensure that the ISI categories are properly ordered in the analysis.

```
/*PROC FORMAT to ensure that the ISI categories are correctly ordered*/
proc format;
value ISIcat
1 = "No clinically significant insomnia"
2 = "Subthreshold insomnia"
3 = "Clinical insomnia (moderate severity)"
4 = "Clinical insomnia (severe)";
run;
```

Denote n_{hij} the number of subjects who are in hsCRP stratum h ($h = 1, \ldots, q$), treatment group i ($i = 1,2$) and have ISI category j ($j = 1, \ldots, r$). Also suppose $\{a_{h1}, \ldots, a_{hr}\}$ is a set of scores associated with each level of the response variable in stratum h. In this

analysis, we use the SAS default "integer scores" such that $\{a_{h1}, \ldots, a_{hr}\} = \{1, \ldots r\}$ for $h = 1, \ldots, q$.

The sum of strata scores for treatment $i = 1$ is

$$f_{+1+} = \sum_{h=1}^{q} \sum_{j=1}^{r} a_{hj} n_{h1j}$$

Under the null hypothesis of no association, f_{+1+} has the expected value and variance

$$E\{f_{+1+}|H_0\} = \sum_{h=1}^{q} n_{h1+}\mu_h = \mu_*,$$

$$V\{f_{+1+}|H_0\} = \sum_{h=1}^{q} \frac{n_{h1+}(n_h - n_{h1+})}{n_h - 1} v_h = v_*,$$

where $\mu_h = \sum_{j=1}^{r}(a_{hj} n_{h+j}/n_h)$ is the finite subpopulation mean and $v_h = \sum_{j=1}^{r}\left(a_{hj} - \mu_h\right)^2 (n_{h+j}/n_h)$ is the variance of scores for stratum h.

The extended CMH test statistic is

$$Q_{SMH} = \frac{(f_{+1+} - \mu_*)^2}{v_*}$$

which has an approximate chi-square distribution with one degree of freedom. The following PROC FREQ implements the test for ISI and the test statistics Q_{SMH} corresponds to "Row mean scores differ" in the output.

```
/*ORDER=FORMATTED is critical*/
proc freq data=ISI order=formatted;
table hsCRP*Treatment*ISIcat / cmh nocum nocol nopercent;
run;
```

Appendix D. Analysis of Continuous Endpoints using Mixed Effect Model Repeat Measurement (MMRM)

The repeated measure analysis will be conducted using mixed model 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 and treatment-by-visit interaction, main stratification factors at randomization 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 and using the method of restricted maximum likelihood (REML).

The MMRM is implemented using the SAS procedure PROC MIXED. The LSMEANS statement provides the least-squares means and corresponding 95% CIs for each treatment group at each visit, as well as the least-squares means, corresponding 95% CIs and p-values for treatment comparisons at each visit. The SAS code example is as follows:

```
proc mixed data=CData;
    class TRTP AVISIT STRATA USUBJID;
    model CHG = TRTP AVISIT BASE STRATA TRTP*AVISIT / DDFM=kr htype = 3;
    repeated AVISIT / subject = USUBJID type=un;
    lsmeans TRTP*AVISIT/ cl diff;
    ods output lsmeans=lsmean_output;
run;
```

Note: The input dataset is per subject per visit. USUBJID is the unique subject number, TRTP denotes the treatment group, STRATA denotes the strata, and AVISIT denotes the visit. CHG denotes the change from baseline value and BASE denotes the baseline value.

Appendix E. Tipping Point Analysis

Tipping Point Analysis for binary endpoint ASAS40 (and ASAS 20)

To assess the robustness of the primary analysis using NRI data handling, tipping point analysis is conducted on the primary endpoint ASAS40. The analysis is conducted using As Observed data handling.

Let M_1 be the number of subjects missing ASAS40 status in the placebo group, and let M_2 be the number of subjects missing ASAS40 status in the upadacitinib group. Let X_1 be the number of subjects imputed as ASAS40 responders out of the M_1 subjects with missing ASAS40 status in the placebo group – the rest are imputed as non-responders. X_1 can take values from 0 to M_1. Similarly define X_2. Given each pair of (X_1, X_2), we can obtain the p-value for the treatment comparison of upadacitinib versus placebo using the combined observed data and imputed data for each treatment group. If one pair of 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 these parameters are called the tipping point.

Tipping Point Analysis for 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 ranked secondary endpoints at the primary time point. The tipping point analyses are two-dimensional, i.e., assumptions about the missing outcomes on the upadacitinib treatment group 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 multiple imputation (MI) under MAR, and then a shift parameter is applied to the imputed values (a different shift parameter may be specified for each treatment group). (The MI imputation is performed upon AO data). This is implemented by PROC MI using the MNAR statement. For a given pair of shift parameters, the SAS code example is as follows:

```
PROC MI DATA=DATA WIDE OUT=OUTMI NIMPUTE=5 SEED=12345
     MINMAXITER=1000;
     CLASS TRTP COVCAT;
     MONOTONE;
     VAR TRTP COVCON COVCAT  WEEK_2 WEEK_4 WEEK_8 WEEK_12;
     MONOTONE REG( WEEK_2  = TRTP COVCON COVCAT );
     MONOTONE REG( WEEK_4  = TRTP WEEK_2 COVCON COVCAT );
     MONOTONE REG( WEEK_8  = TRTP WEEK_2 WEEK_4 COVCON COVCAT );
     MONOTONE REG( WEEK_12 = TRTP WEEK_2 WEEK_4 WEEK_8 COVCON COVCAT );
     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 the same as that used for MI analysis. The macro variables SJ1 and SJ2 denote the shift parameters for the placebo group and the upadacitinib group respectively.

In cases where the shifted values are smaller than the minimum or larger than maximum value of the endpoints, (i.e., out of range), the minimum or maximum value is used in further analysis steps. For each pair of shift parameters, the same analysis as the primary MI analysis is then conducted to obtain the p-values for the upadacitinib treatment group versus the placebo group. 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.

Appendix F. Geographic Region

Below table lists the countries considered for each geographic region.

Geographic Region	Countries
Asia	Japan, South Korea
Eastern Europe	Croatia, Czech Republic, Hungary, Poland
North America	Canada, United States
Other	Australia, New Zealand
Western Europe	Belgium, Denmark, Finland, France, Germany, Netherlands, Portugal, Spain Sweden, United Kingdom

Appendix G. Low Dose CT Imaging

Per study protocol, subjects enrolled into Study M16-098 can choose to enroll into a sub-study in which low dose CT imaging of the spine is performed. The analysis population for low dose CT includes all the subjects who consented to the sub-study, were randomized, with at least one dose of study drug, and with at least one measurement of low dose CT. Analysis will be conducted using As Observed (AO) cases.

Each subject enrolled in the low dose CT sub-study should be measured at baseline, Week 52 and Week 104. The summary of the low dose CT score change from baseline to Week 52 or Week 104 will be summarized by treatment group sequence as defined in section 9.1.1 and the overall group. The descriptive statistics min, max, mean, SD, and the 95% CI of mean for each group at each visit will be provided. The CT scans are evaluated for the full spine (cervical, thoracic, and lumbar spine) by the central image reading and the scoring of the low dose CT Syndesmophyte is described below.

At each timepoint, the reviewer will score a total of 23 vertebral units (VUs), from C2-S1, in the coronal and sagittal planes. In the sagittal view, the anterior and posterior rim of a vertebra can be assessed whereas the left and right rim can be assessed in the coronal view.

By scoring images in the sagittal and coronal planes, 8 quadrants are assessed per VU (i.e., per VU lower and upper borders: coronal - left rim, coronal - right rim, sagittal - anterior rim, and sagittal - posterior rim). With 8 quadrants per vertebral unit, 23 VUs per spine, and a max score of 3 per quadrant, the maximum possible score for a subject is 552 (ACR, 2016, de Bruin, 2017).

A syndesmophyte is defined as a bony spur arising from the vertebral body close to the vertebral endplate in a vertical configuration with an angle of ≥ 45-degrees in relation to the horizontal endplate. For each quadrant, the reviewer will enter a syndesmophyte score as follows:

0 = no syndesmophyte present

1 = a syndesmophyte is present that does not reach 50% of the intervertebral disc space (IDS)

2 = the syndesmophyte reaches or crosses 50% of the IDS

3 = the syndesmophyte bridges the IDS

A value of 3 is, per definition, scored in both quadrants on opposite sides of the intervertebral disc space. If a VU is not visible (e.g., poor film quality, missing imaging,) at the timepoint, the individual VU will be coded as 'Not Visible' (N). If images at the timepoint show a VU with surgical fusion, then the VU will be scored 'Surgically Modified' (S). No numerical values will be provided for these locations and will be considered as missing.

If a score at any location (vertebral unit) is missing, the method described below will be used for deriving syndesmophyte score.

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